Development of a protocol for environmental PCR of 18S rDNA V4 region and diversity survey of freshwater protists using 454 Titanium pyrosequencing

Dan Kristofer Ree

Master of science thesis
Department of Biology
Microbial Evolution Research Group (MERG)

UNIVERSITY OF OSLO
June 8th 2010

Acknowledgements

The laboratory and bioinformatical part was carried at the facilities of the Microbial Evolution Research Group (MERG) from October 2008 to June 2010.

First of all I would like to thank Dag Klaveness for giving me the opportunity of doing my thesis on such a fascinating subject and for the support, motivation, help talks and field trips provided. Kamran Shalchian-Tabrizi and Jon Bråte for always being available for anything I could ask for, but most of all their tireless work on making this thesis what it is. I could not have done this without you three.

I would also like to thank all the "guys at the office" for their help with different aspects of the work associated with this thesis, and their good mood

Anders for all the laughs and for sharing your great knowledge of many things.

Surendra for always smiling and being able to do all the things with a computer that I cannot.

Sen for always being willing to help, support or laugh.

And Russell for sharing what he knows and being positive.

I must also thank my family and friends for putting up with me these last weeks, I know I have been absent most of the time.

Last but not least, I would like to thank my beautiful wife, for waiting for me, once again, and doing such a great job of carrying our first child.

Dan Kristofer Ree, Blindern, June 2008.

Table of Contents

Introduction

Studies of microbial diversity have traditionally been largely dependent on the proper identification of species in a light microscope. However, as many microbial species lack features that are distinguishable in the microscope only a small number of microbial species had been described well into the 1900's. As many species lack distinguishing morphological features they were often mistakenly lumped together into the same taxa so that most of them were found distributed worldwide, which led to the theory that "everything is everywhere, but the environment selects" (Bass-Becking 1934, de Wit and de Bouvier 2006). All eukaryotes can be divided into only a few supergroups as shown in figure 1 and the majority of this diversity is to be found among the single celled eukaryotic life forms (protists). This means that a large part of the protist diversity was undiscovered either because of misidentification or because they were too small to be properly study under the light microscope.

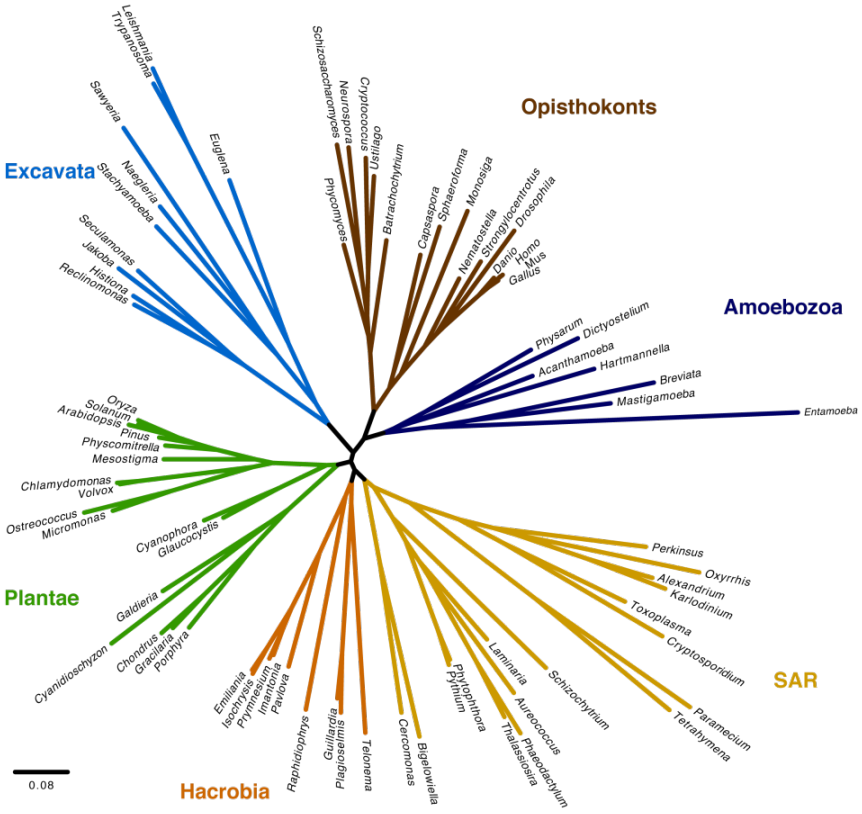


Figure 1. Global eukaryotic phylogeny showing the different supergroups (tree inferred from 137 gene sequences. Shalchian-Tabrizi et al unpublished).

As science uncovered more of the secrets of DNA it became apparent that there was variation in the genetic material between species and that this might be used to study taxonomical relationships. By the 1960s' this was becoming a more recognized fact and a method describing the use of DNA to investigate taxonomic and evolutionary relationships between species was published in 1963 (McCarthy and Bolton 1963). With increased knowledge of genetics it was possible to sequence the information contained in DNA and this was first applied to biodiversity studies in 1985, when a group of researchers published an article describing an approach using genetic material to identify individual bacterial organisms in the population based on ribosomal DNA (rDNA) sequences, giving rise to modern metagenomics and environmental sequencing (Pace et al 1986).

Environmental sequencing proved very efficient and during the 1990's environmental sequencing was used to demonstrate the presence of Archaea in most habitats and to show that the diversity of this group was much larger than previously known (De Long and Pace 2001). In 2001 this method was applied to study the diversity of protists in environmental samples from oceanic regions and it was immediately clear that the diversity of protists in these habitats was much greater than what was previously identified by culturing (Lopez-Garcia 2001; Moon-van der Stay 2001). Many new and previously unknown groups was uncovered, such as the groups referred to as Marine stramenopiles and Marine alveolates, both large and diverse groups of heterotrophic flagellates found worldwide and which likely plays important roles in the aquatic ecosystems (Lopez-Garcia et al 2001; Moon-van der Stay et al 2001). Since then, environmental sequencing has greatly increased what we know of the protist diversity and shown that there still remains a large amount of undescribed species that have gone undetected by microscopy-based investigations.

The discovery of the extent of the "hidden biosphere" of microbes led to a demand for a cheaper and less time consuming way to generate sequence data than Sanger-based sequencing. In 2005 a method combining the use of emulsion PCR and pyrophosphate-based sequencing, in several picoliter-sized wells on a PicoTiterPlate was published (Marguiles et al 2005). This made it possible to produce large amounts of sequence data without the need of an expensive and time consuming bacterial cloning step required by Sanger sequencing. The pyrosequencing technology was made commercially available with the development of the 454 GS FLX pyrosequencing platform from 454 Life sciences, and the first version was able to produce sequences up to 250 base pairs (bp) in length. The length of sequenced reads has been increased to about 500 bp with the new Titanium upgrade of the 454 pyrosequencing

platform. The studies referred to earlier have targeted parts of the small ribosomal subunit gene, 16S rDNA in bacteria and 18S rDNA in eukaryotes. These genes were chosen primarily for two reasons; they are present in all known living cellular organisms and due to different selection pressures across the gene consist of highly variable areas separated by conserved regions. This makes these two genes useful in phylogenetic studies of both distant and closely related species.

The first surveys of eukaryote diversity using the 454 pyrosequencing targeted the variable V9 region (150 bp) of the 18S rDNA (Amaral-Zettler et al 2009). The increased output length provided by the latest Titanium upgrade makes it possible to target longer regions to use as genetic markers to identify individual organisms. And with more information contained in each sequence, species identification and phylogenetic analyses becomes more robust. With a length of up to 500 bp, the V4 region is one of the largest and most variable of the regions in the 18s rDNA (Nickrent et al 1991) and is flanked by conserved regions suitable for design of universal primers that should in theory amplify a wide range of eukaryotes, making it an ideal target for biodiversity studies.

Although a few environmental studies of the protist diversity have been done on freshwater samples, but never on freshwater sediments, marine waters are by far the most investigated. Freshwater habitats are much more heterogeneous and diverse than oceanic habitats so the protist diversity in freshwater is most probably vastly underestimated. The few studies that have been performed supports in that they all have revealed a large and unknown diversity especially among the pico-sized and heterotrophic protist (Slapeta et al 2005; Richards et al 2005; Lefèvre et al 2008).

Aims

As there has been only few diversity surveys in freshwater and because an increasing amount of diversity is being uncovered in freshwater habitats, it is a great need to develop more efficient molecular methods that can utilize the advances in technology development. Based on the above-mentioned themes, I have therefore defined two complementary main aims in my master thesis.

1. To develop a protocol for environmental PCR studies of eukaryotes suitable for the new 454 Titanium pyrosequencing technology.

2. To study the eukaryote diversity in sediment samples from a Norwegian freshwater lake, Lake Finsevatn.

Central to these main aims are several questions that I wish to address:

- Can we develop universal eukaryote primers that are suitable for the 454 GS FLX Titanium pyrosequencing technology?
- Is the V4 region of the 18S rDNA gene suitable as a genetic marker for diversity surveys and adequate as target for environmental PCR and 454 Titanium sequencing?
- Will large-scale sequencing reveal species of protists that have not been observed in freshwater?
- Can we identify larger groups of new protists in our freshwater DNA sample?

About my contributions in this thesis and the attached papers

The first sequencing run performed during my master thesis resulted in two articles: one article accepted in the ISMEJ (Paper 1) and a manuscript submitted to Biology Letters (Paper 2). As several people have been involved in the work of with these articles and this thesis, it is important for me to clarify my contributions to these papers (see appendices).: I took part in the development of the presented sequencing protocol and design of the universal primers as well as some of the lab work and phylogenetic analyses associated with the first round of

sequencing together with Jon Bråte, as part of my master program. The protocol we made and the new diversity we identified resulted in two multiauthored papers attached as an appendix.

In the attached paper "Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental DNA" (Bråte et al 2010; Paper 1) I participated in the bioinformatic and phylogenetic analyses involved in determining the relationship of our reads to the group in question. The rest of the analyses and writing of the manuscript were done by the other authors. In the second attached manuscript "Pyrosequencing of 18S rDNA reveals unknown haptophytes in freshwater and an ancient marine-freshwater divergence in Prymnesiophyceae" (Shalchian-Tabrizi et al. unpublished; Paper 2), I designed 3 class specific primers targeting a subgroup of Haptophyceae that are also presented in this thesis. These primers were used by others in a study of freshwater haptophyte diversity.

After the submission of these papers I repeated the entire lab protocol as part of the second round of sequencing. I also performed the bioinformatic and phylogenetic analyses associated with the second run of sequencing. The results obtained from this work are presented in this thesis and has not been included in Paper 1 and 2. This thesis is a presentation of the developed method and primers, and summarizes all methods and results related to the two sequencing runs performed on the sediments from Lake Finsevatn.

Methods

Sampling and DNA extraction

The sample site, Lake Finsevatn, is a high alpine (1215m a.s.l.), oligo-mesotrofic lake with a monomictic circulation. It is situated in the northwest part of the Hardangervidda at location 60˚36' N – 7˚30 E in Ulvik county, Norway. Diversity in the lake has been sampled twice, first by the Norwegian Institute for Water Research (NIVA) in the summer of 1985 (Aanes et al 1987) and from the summer of 2003 to the summer of 2004 (Hagnar 2006). Both studies were based on microscopy investigations using morphological characters for identification of species. Due to the difficulty of assigning organisms to groups by morphological characters, almost 16% of the total biomass was considered unassignable (Hagnar 2006).

A sediment sample was taken on March 5th 2009 using a simple gravity corer at 18 m depth and filtered through a Millipore Durapore 0,22 µm membrane filter. DNA was extracted using the Power Soil DNA kit (MoBio, Carslbad, CA, USA) following the manufacturers' protocol, except for the bead beating that was done on a mixer mill (MM 301, Retsch GmbH & CO, Haan, Germany) with a frequency of 12 Hz for 10 minutes. Four parallel isolations were made, each from a quarter of the filter. After extraction samples were measured for content of DNA on Nanodrop 1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA).

Designing universal primers

Primers were designed in an alignment of the 18S rDNA gene containing a wide selection of eukaryote species from Berney et al (2004) and the V4 region was identified using the secondary structure model in figure 2. Two preserved areas were found demarcating the target region. We decided to use an already published universal eukaryote primer 3NDf (5'-GGCAAGTCTGGTG CCAG-3') that matched the conserved area upstream of the V4 region (Cavalier-Smith et al 2009). Three reverse primers were designed that annealed to the conserved area downstream of the V4 (shown in table 1). The combination of these primers would produce amplicons up to 500bp in length, which is recommended by 454 Life

Sciences. Characters in the primer sequence containing more than one nucleotide were assigned as degenerated nucleotides. Ambiguity codes were used in cases where no single common nucleotide was shared among all eukaryote groups. All primers were assessed using Basic Local Alignment Tool (BLAST) against the NCBI (National Centre of Biotechnology Information) nucleotide non-redundant (nr) database and ProbeCheck (Loy et al 2008) against the ARB-SILVA database to check the generality of the primers. PRIMER3 (Rozen and Skaletsky 1998) was also used to check primer characteristics such as the theoretical melting temperature (Tm) and the content of guanine (G) and cytosine (C).

Table 1. The designed universal eukaryote primers are shown with theoretical melting temperature (Tm) content of GC and length. Ambiguities are shown in parentheses.

Primer name	Sequence 5'-3'	Tm (^{o}C)	GC (%)	Length (bp)
V4_euk_R1	GACTACGACGGTATCT(AG)ATC(AG)TCTTCG	65,0	48,1	27
V4_euk_R2	ACGGTATCT(AG)ATC(AG)TCTTCG	55,3	45,0	20
V4_euk_R3	CCGTCAATTCCTTTAAGTTTCAG	57,1	39,1	23

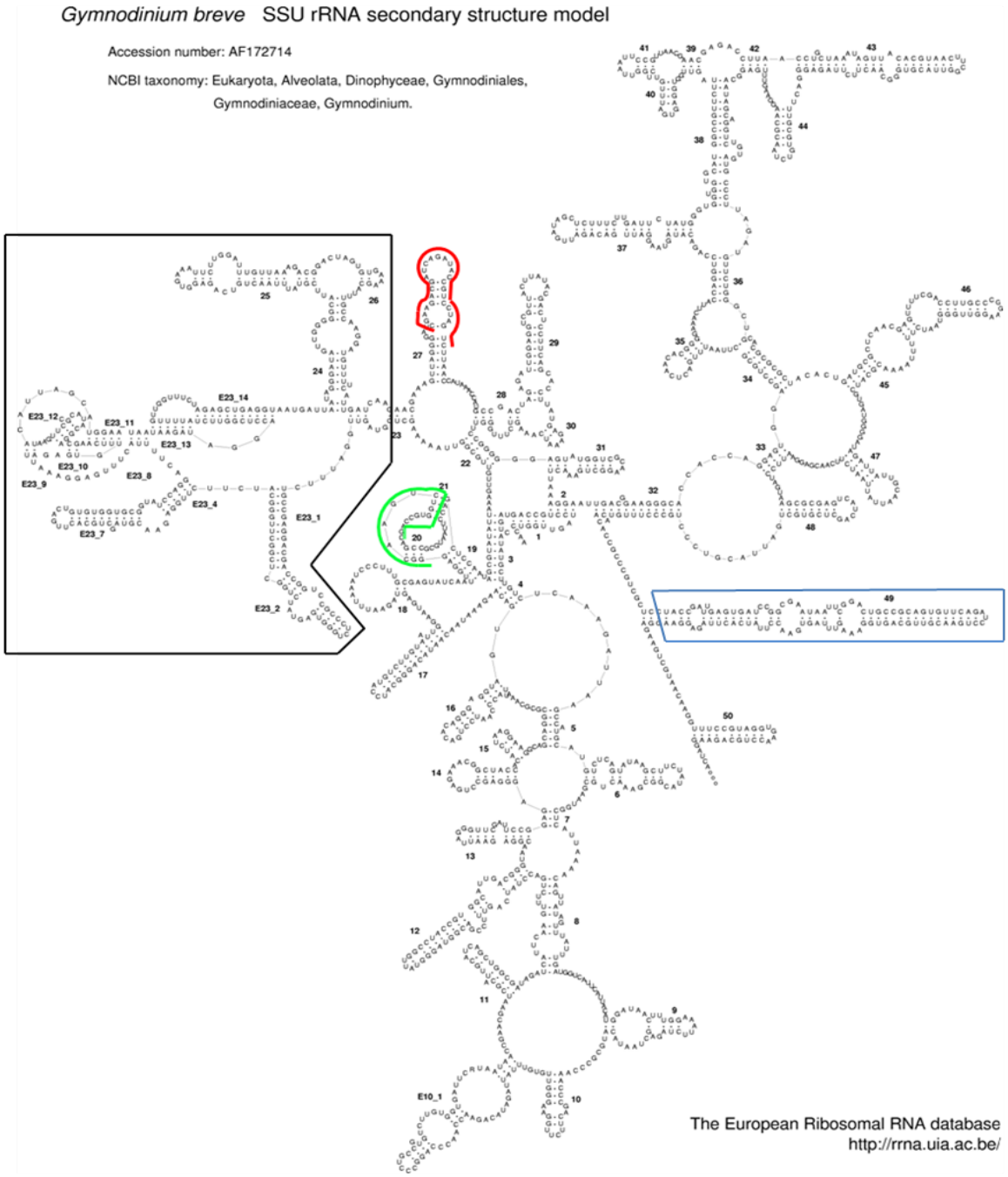


Figure 2. 18S rDNA secondary structure model of a eukaryote (*Gymnodinium breve* Accenssion number: AF172714). Region V9 is marked by a blue box and V4 is marked by a black box. Forward primer 3NDf indicated in green and reverse primers V4_euk_R1 and V4_euk_R2 indicated in red (figure modified from The European Ribosomal RNA database [http://bioinformatics.psb.ugent.be/webtools/rRNA/]).

Amplicons generated by our reverse primers were verified and checked if amplicons were within the desired target range (up to 500 bp), by visualization using electrophoresis on a 1%

agarose gel. Reverse primers were tested individually against the forward primer 3NDf and amplicons were generated as described below.

The PCR was run in 25µl reactions containing 1µl of 1:1, 1:10 diluted template, 2,5µl of 1x HotMaster (5 PRIME, Hamburg, Germany) Taq Buffer with 2,5mM Mg^{2+}, 0,25µl 1,5UHotMaster Taq DNA Polymerase, 2,5µl dNTPs (2µM), 0,5µl of both forward and reverse primers and 17,75µl of Milli-Q (Mq) water.

A positive control was included using DNA isolated from a cryptomonad culture (donated by Dag Klaveness), and a negative no control was included to check for contamination of the reagents and assessment of the PCR reaction.

The amplification was done using an initial denaturation temperature of 94^{o}C for 2 minutes then 35 cycles of 94^{o}C for 30 seconds, 58^{o}C for 30 seconds and 68^{o}C for 2 minutes and 30 seconds, the final extension was 68^{o}C for 7 minutes. PCR was done using an Eppendorf Master Cycler ep (Eppendorf, Hamburg, Germany).

The PCR products were then cleaned using Wizard SV Gel and PCR Clean-Up System (Promega Corporation, Madison, WI, USA) and cloned using TOPO TA cloning kit (Invitrogen, Carlsbad, CA USA) following the manufactures instructions. 10 white colonies were then checked for positive inserts by a second amplification run using primers TopoF and TopoR in an equal reaction mix as described above, except the use of 1x DyNAzyme (Finnzymes Oy, Espoo, Finland) buffer (F-511) and DyNAzyme II DNA polymerase (1,5U). The latter amplification was done with a denaturation temperature of 95^{o}C for 2 minutes, then 35 cycles of 95^{o}C for 20 seconds, 60^{o}C for 20 seconds and 72^{o}C for 50 seconds, and a final extension of 72^{o}C for 7 minutes.

Electrophoresis was done by staining the samples with 6X DNA Loading Dye (Fermentas International, Burlington, Canada) and then loading them together with a with a FastRuler Low Range DNA Ladder (50-1500bp) (Fermentas International, Burlington, Canada) on a 1% agarose gel stained with either ethidium bromide or SYBR Safe nucleic acid stain (Invitrogen,Carlsbad, CA USA). Visual inspection of the PCR products was done by electrophoresis on a 1% agarose gel and left on 90V for 45 minutes and visualized using UV illumination using Geneflash Gelvue transilluminator and GeneFlash gel documentation system (Syngene, Cambridge UK).

Designing genus specific primers

Primers specific to the Haptophyta class Pavlovophyceae were also designed. These were positioned in the middle of the 18S rDNA V4 region and were used together with universal eukaryotic forward and reverse primers matching either either of the two flanking ends of the 18S rDNA gene. This was done by adding sequences of known Pavlovophyceae origin to a general eukaryotic alignment (from Berney et al 2004) along with reads identified as Pavlovophyceae from Lake Finsevatn. Areas of the V4 region that were specifically conserved among only members of Pavlovophyceae were located and three primers were designed that were of different lengths. Primers were then designed for these areas following the same method as described for the universal primers. Each of the primers was made in both forward and reverse direction (reverse complemented).

The primers designed are shown in table 2 and were tested by the same methods as described for the universal primers.

Table 2. The designed Pavlovophyceae specific primers with melting temperature (Tm), content of GC and primer lengths are shown.

Primer name	Sequence 5'-3'	Tm (^{o}C)	GC(%)	Length (bp)
Pavlova_v4_F1	TCGTATTCCGTAGAGAGAGGT	57,9	47,6	21
Pavlova_v4_R1	ACCTCTCTCTACGGAATACGA	57.9	47,6	21
Pavlova_v4_F2	GTGAAATTCTTAGACCCACGGA	58,4	45,5	22
Pavlova_v4_R2	TTCCGTGGGTCTAAGAATTTCA	56,5	40,9	22
Pavlova_v4_F3	ATTGTATGTGTTAGCATGGGATAATGGA	60,7	35,7	28
Pavlova_v4_R3	TCCATTATCCCATGCTAACACATACAAT	60,7	35,7	28

Optimalization of PCR amplification

Sequencing on a 454 GS FLX Titanium machine (454 Life Sciences, CT, USA) requires the adding of adaptor sequences for emulsion PCR and Multiplex Identifiers (MIDs) to allow for parallel sequencing of several individual samples. Adaptor sequences and MIDs were added to the designed eukaryotic primers, together making up the composite primers shown in table 3.

The composite primers were significantly longer than the original primers and the requirements changed as such, with the Tm climbing to more than 75°C and a PCR program had to be adapted to suit these new requirements. After Optimalization of the PCR protocol the original 35 cycles were divided into two separate cycles with different annealing temperature to allow for these new requirements. Low temperatures allow for annealing of only the V4 specific part of the composite primers to bind without the tag and mid attaching during the first cycles. After several runs with different annealing temperatures and temperature gradients, the optimal temperature was found to be 60°C for the first cycle for the second cycle the temperature was set at 65°C.

Table 3. Showing composite primers, with length, GC content and Tm. Adaptor sequences are shown in bold, MIDs are in italic.

Primer name	Sequence 5'-3'	Length (bp)	GC (%)	Tm °C
454_V4_euk_F1	**CCATCTCATCCCTGCGTGTCTCCGACTCAG** *CGTGTCTCTA*GGCAAGTCTGGTGCCAG	57	57,9	>75
454_V4_euk_R1	**CCTATCCCCTGTGTGCCTTGGCAGTCTCAG** GACTACGACGGTATCT(AG)ATC(AG)TCTTCG	57	54,4	>75
454_V4_euk_R2	**CCTATCCCCTGTGTGCCTTGGCAGTCTCAG** ACGGTATCT(AG)ATC(AG)TCTTCG	50	54,0	>75

Amplicon generation

In order to generate enough DNA template for the 454 sequencing, we first amplified the V4 region with the universal primers (table 1) The produced amplicons were then applied as template for a second amplification step using the composite primers (table 3). Parallels were used in all amplifications to reduce random PCR bias and further increase DNA content by pooling after finished amplification (Engelbrektson et al 2010). The steps in this process are described below and summarized in figure 3.

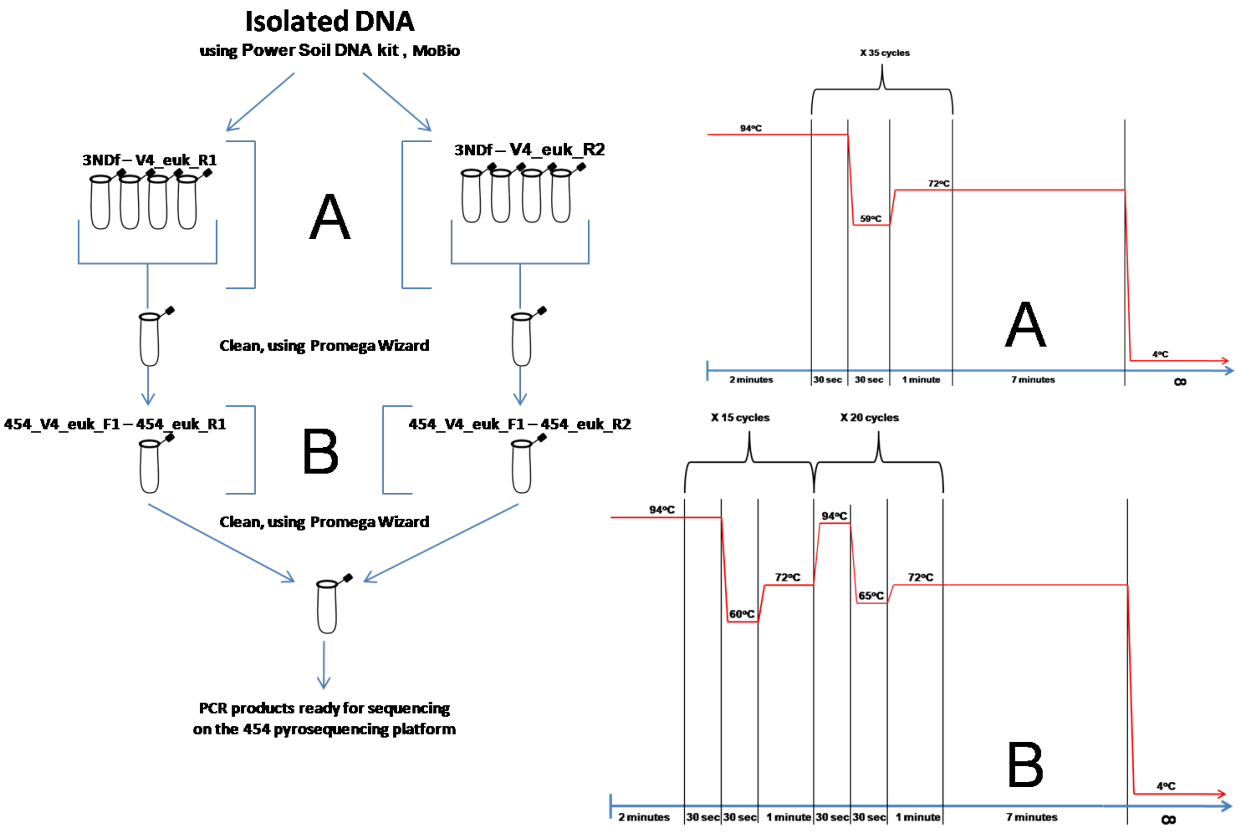


Figure 3. Summary of the amplification protocol developed for the preparation of PCR products for 454 Titanium pyrosequencing on the right side, applied PCR programs shown on the right. A: first amplification reaction using the designed eukaryote primers, B: second amplification reaction using the composite primers.

The first amplification step was done using forward primer 3NDf and reverse primers V4_euk_R1 and V4_euk_R2. PCR was run in 25µl reactions containing 1µl of 1:10 diluted template, 2,5µl Fermentas 1x Dream Taq buffer with2,5 mM Mg^{2+} (Fermentas International, Burlington, Canada), 0,125µl Fermentas 0,6U Dream Taq polymerase, 2,5µl dNTPs (2µM) and 0,5µl of both forward and reverse primers (10 µM). Mq water was added to a total of 25µl. The PCR was done using an Eppendorf Master Cycler ep (Eppendorf, Hamburg, Germany), and 4 parallels were done for both primer combinations. Each of the parallels were pooled and cleaned using Wizard SV Gel and PCR Clean-Up System (Promega, Madison, WI, USA), to remove unincorporated primers and purify the DNA. The PCR program for the first round was: an initial denaturation of 94°C for 2 minutes followed by 35 cycles of 94°C for 30 seconds, 59°C for 30 seconds and 72°C for 1 minute, with a final extension step of 72°C for 7 minutes.

In the second step the composite primers were used to add primer for emulsion PCR (emPCR) and multiplexing tags (MIDs) to the initial amplicons. The PCR program for the second

amplification run was: an initial denaturation of 94°C 2 minutes, then 15°C cycles of 94°C for 30 seconds, 60°C for 30 seconds and 72°C for 1 minute followed by 20 cycles of 94°C for 30 seconds, 65°C for 30 seconds and 72°C for 1 min, and a final extension of 72°C for 7 minutes.

After each amplification step the amplicon lengths was verified by gel electrophoresis, as described earlier.

Pyrosequencing, filtering and clustering of reads

Amplicons were sequenced by standard procedures on a GS FLX Titanium machine following instructions from manufacturer (454 Life Sciences, Brandford, CT, USA).

Qualtiy assessment, removal of quality sequences and clustering was done using Phylosity, a pipeline that filters and removes unwanted reads, and clusters sequences at parameters set by user (Kumar et al 2010). Unwanted reads could have low quality due to primer or tag error, short sequences, degenerate characters (N) or error in repeated genetic characters (homopolymers), which is one of the most common sources of error in pyrosequencing (Marguiles et al 2005).

Phylosity parameters were set as follows, first sequences shorter than 300 bp and containing the degenerate code N were removed and homopolymers longer than 8 bp were collapsed to eight. Sequences were clustered using single linkage clustering, to reduce possible overestimates of diversity created by pyrosequencing or PCR errors. Clustering criteria were set at 98% similarity and 75% coverage and the longest sequences were retained and then used as query sequences using Basic Local Alignment Tool (BLAST) against the NCBI (National Centre of Biotechnology Information, USA) nucleotide non-redundant collection (nr) database on the Bioportal (www.bioportal.uio.no).

Data mining and identification of cryptomonad sequences

Results from BLAST were investigated using MetaGenome Analyzer (MEGAN) where reads are assigned to described eukaryotic groups based on the BLAST query output (Huson et al 2007). Last common ancestor (LCA) parameters were set at score = 1 and bit score = 300, meaning that the top hit from the BLAST output was used if the bit score was over 300. Bit scores are derived from a substitution matrix for each position in the alignment. Reads with

BLAST hits of unknown origin, were placed in the "not assigned" group and reads that generated no hits in the database were placed in the "no hits" group. The "not assigned" group was manually checked for sequences of possible Cryptophyceae origin by inspecting the BLAST output for each sequence.

To confirm relationship to cryptophytes, all sequences extracted from the BLAST output, either manually from the "not assigned" group or directly identified by MEGAN, were inserted into the global eukaryote alignment (Berney et al 2004) hereafter referred to as Al 1. Sequences were aligned using Mesquite v 2.72 (Maddison and Maddison 2009) and the Opal module set at default settings (Wheeler and Kececioglu 2007). The alignment was then manually edited and ambiguously aligned regions were excluded from further analysis.

Phylogenetic trees were created by maximum likelihood analyses (as described below). Studying the generated trees we excluded reads not placed among cryptophytes. The remaining sequences were then included, using Mesquite v 2.72, into a Cryptophyceae specific alignment (Al 2) to allow for the inclusion of more unambiguously aligned characters (Cryptophyceae alignment provided by Jon Bråte).

Phylogenetic analysis

Alignments were uploaded to the Bioportal (www.bioportal.uio.no) and all computational analysis was performed there.

Using Modeltest (Posada and Crandall 1998) the General Time Reversible model (GTR) with a gamma distribution rate of variation across the sites (G) and a proportion of invariable sites (I) was inferred as the most optimalt evolutionary model for the dataset according to hierarchical likelihood test, and was used for all applicable analysis.

Maximum likelihood (ML) analyses were done in RaxML v.7.0.4 (Stamatakis 2006). Due to the large size of the dataset a heuristic tree search with rapid hill climbing from a random starting tree was performed. 4 rate categories were calculated and 100 independent inferences run, from which we selected the topology with the highest likelihood.

Bootstrapping was done to statistically test the ML topology. The analyses were executed with the same model and parameters as for the original ML analysis, with100 pseudo replicates were produced, from which a majority rule consensus tree was constructed.

The entire bioinformatical workflow used to process the 454 pyrosequencing reads is visualized in figure 4.

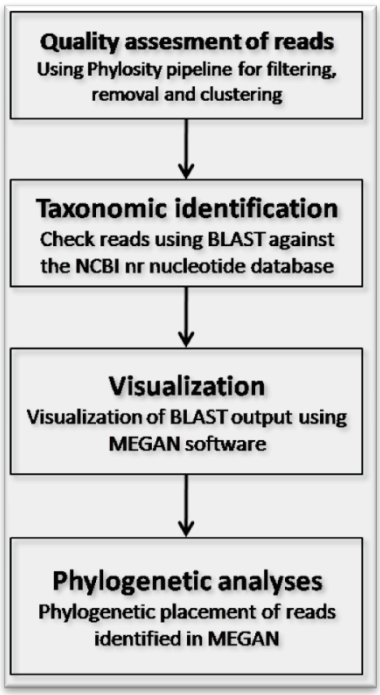


Figure 4. Summarizing the main bioinformatical steps involved in the analyses of the 454.

Results and discussion

The developed protocol for PCR of the V4 region of 18S rDNA successfully amplified DNA from environmental samples

One of the main aims of this thesis was to develop a protocol for environmental PCR of the V4 region of the 18S rDNA gene that is suitable for diversity studies of eukaryotes, and optimized for the newest 454 pyrosequenecing technology. In our approach for developing such a protocol we designed three PCR primers that in theory should amplify most of the eukaryote diversity (i.e. universal primers). All three primers were matched to a conserved region at the 3' end of the V4 region (i.e. reverse primers), and were used against an already published primer at the 5' end of the region (i.e. forward primer).

The DNA template for testing the primers, which was gained from sediments of Finsevatn had an average estimated DNA concentration of 7,8 ng/µl for the 4 parallel DNA isolations.

Using this DNA, all three designed reverse primers generated amplicons by PCR in the desired target size range of about 500 bp in combination with the forward primer 3NDf (Cavalier-Smith et al 2009). The primer V4_euk_R3 however also produced amplicons that were longer than this as is shown by the dual bands in the lanes assigned to this primer in figure 5. The primer V4_euk _R3 therefore seemed less specific than the two others and hence excluded from further analyses. Nevertheless, despite relatively low DNA concentrations, we were able to generate amplicons in all PCR reactions. After the first round of PCR the average DNA concentration of amplicons was 34,8 ng/µl and after the second round of amplification using the composite primers (table 3) the average concentration was 54,4 ng/µl of DNA. Verification of all generated amplicons by UV illumination on an agarose gel, after each PCR, confirms that amplicons from all reactions are in the target range of the V4 region of the 18S gene (figure 5).

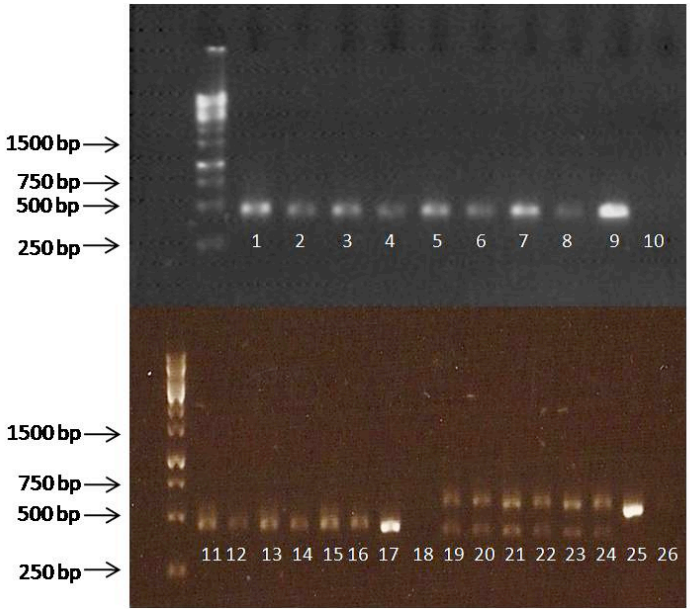


Figure 5. Verification of PCR products amplified by the three designed universal V4 18S rDNA primers (**table 1**), on a 1% agarose gel after electrophoresis. Ladder used: Fermentas 1kb ladder (Fermentas International, Burlington, Canada). Lane 1-10: PCR product amplified using primer combination 3NDf (Cavalier-Smith) and V4_euk_R1. Templates used are 1 and 1:10 dilutions of 4 aliquots from the DNA isolated from Lake Finsevatn. Lane 11-18: PCR product amplified using primer combination 3NDf (Cavalier-Smith) and V4_euk_R2. Lane 19-26: PCR products amplified using primer combination 3NDf (Cavalier-Smith). Both these primer combinations use 1 and 1:10 dilution of 3 different aliquots of Lake Finsevatn DNA. Lane 9, 17 and 25 are positive controls using template DNA isolated from a Cryptophyceae culture. Lane 10, 18 and 26 are negative controls.

Amplicon sequencing of V4 region using 454 GS FLX Titanium sequencing technology

In order to assess the quality of the amplicons generated by our designed primers, we sequenced a small sample with the 454 GS FLX Titanium instrument. The sequencing resulted in 9937 reads with an average length of 237 bp. The total length distribution of the sequenced reads is shown in figure 6. From this graph we can see two distinct peaks in the length distribution, one at about 30 bp and one about 450 bp. The large amounts of reads of about 30 bp in length are probably primer dimers that remained in the PCR products after the applied cleaning step. This indicates that the PCR amplification and cleanup of PCR products can be optimized further to avoid as many short reads in future work. In contrast, the peak of read lengths around 450 bp suggests that we have been able to amplify the V4 region of the 18S gene.

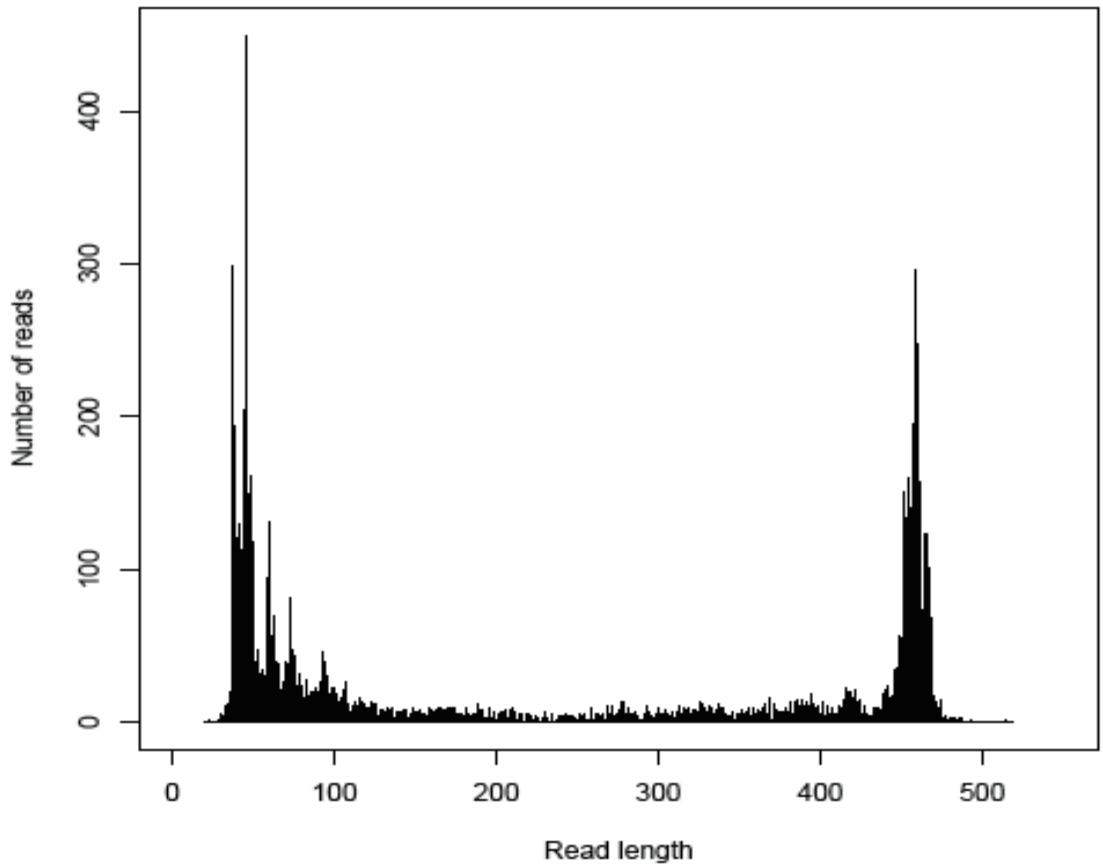


Figure 6. The length distribution of reads sequenced in the first round of pyrosequencing on the 454 GS FLX Titanium pyrosequencing platform.

BLAST searches confirmed the amplification of the V4 region and identified new protist species

After removal of low quality sequences we were left with 4043 reads, implying that more than 50% of the reads were rendered unsuitable for further analyses. The main reason for the considerable reduction of useful reads is the large amount of primer dimer in our sequences. Clustering of the remainging reads resulted in 578 clusters.

Using these 578 clusters as query sequences in BLAST search against the NCBInr database confirmed that the amplicon we produced was the V4 region of the 18S gene. In fact, we could not identify any other type of sequences in our data, suggesting that the primers are highly specific for the 18S gene and hence useful for diversity surveys of the eukaryote diversity by environmental PCR.

A pilot biodiversity survey of Lake Finsevatn revealed new diversity of protists

The second main aim of this thesis was to investigate the unknown protist diversity that is to be found in freshwater sediment. In order to reveal the diversity of the eukaryotes we had sequenced, the BLAST results was parced with the MEGAN program (**figure 7**) and shows sequences spanning over almost all supergroups of eukaryotes (Burki et al 2007). In fact all eukaryotic supergroups except Excavata were clearly identified, including all groups described by earlier microscopy investigations in Lake Finsevatn (Aanes et al 1987; Hagnar 2006). Furthermore, several groups of protists that never, or very rarely, have been observed in freshwater habitats could be identified, such as Telonemia, *Ancyromonas,* Pavlovales and 4 unclassified alveolates (figure 11). Altogether, these results demonstrate that the developed protocol and primers for PCR of the V4 18S rDNA region is able to uncover a wide range of eukaryotes and also sensitive enough to detect even small cell numbers.

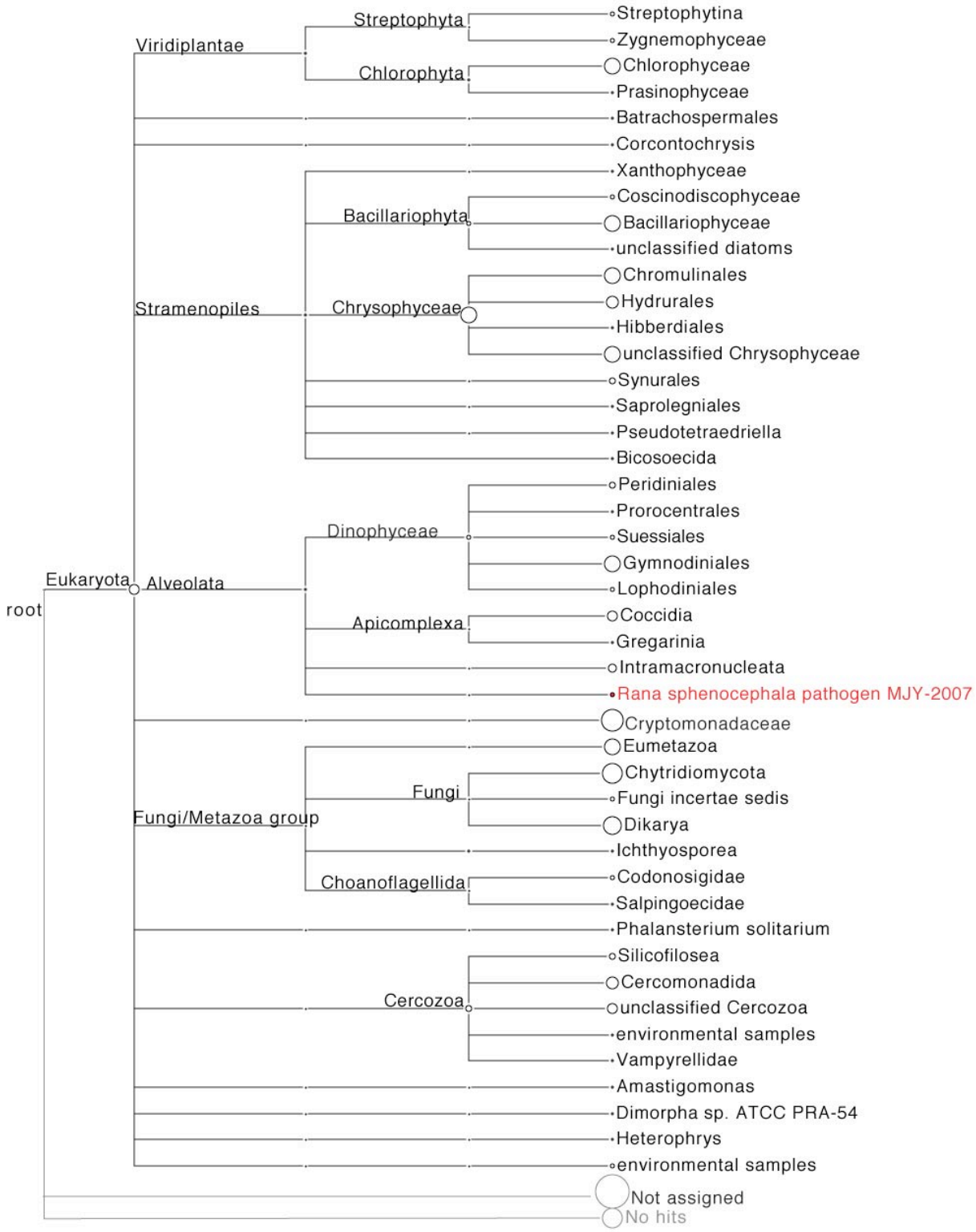


Figure 7. BLAST results from pyrosequencing run 1, visualized in MEGAN. The *Rana spenochephala* pathogen is highlighted in red (for discussion see text).

The PCR protocol detected novel freshwater groups

New studies are now starting to uncover a large diversity of parasitic protists and one of the most diverse groups containing parasites are Alveolata. This group contains, among others, several parasitic lineages related to Dinophyceae, such as *Perkinsus* and *Parvilucifera* (class Perkinsea), groups that never before have been described in freshwater. In the BLAST query results from sequencing run 1, several reads were related to the *Rana spenochephala* pathogen (figure 7). This caught out interest as previous works has shown the presence of a potential unknown freshwater diversity related to this parasite (Bråte 2008). The *Rana sphenocephala* pathogen is a unicellular eukaryote related to Perkinsea and has been shown to infect and cause mass killings among Leopard frogs in the US (Davis et al 2007). Out of the reads identified in the BLAST search, 13 of the sequences generated were in the global phylogeny firmly placed as a distinct group together with *Perkinsus* and *Parvilucifera* (figure 8). These sequences imply that there is a large and unknown diversity of *Perkinsus*-like species in freshwater, even in high mountain lakes. This is confirmed by other sequences that were identified in GenBank along the analyses of our 454 reads; in fact Perkinsea could be divided into 17 distinct groups, which the majority is from freshwater (figure 2 in Paper 1 attached). One of the GenBank sequences originally obtained from a Chinese lake was placed in the Finsevatn group, showing that the group we detected may have a widespread and perhaps global distribution (figure 8). Hence, application of efficient molecular methods makes it possible to detect parasites that have not been detected before by microscopy of freshwater samples. This work on Perkinsea is assembled into my first publication here presented as Paper 1 (Appendix 1).

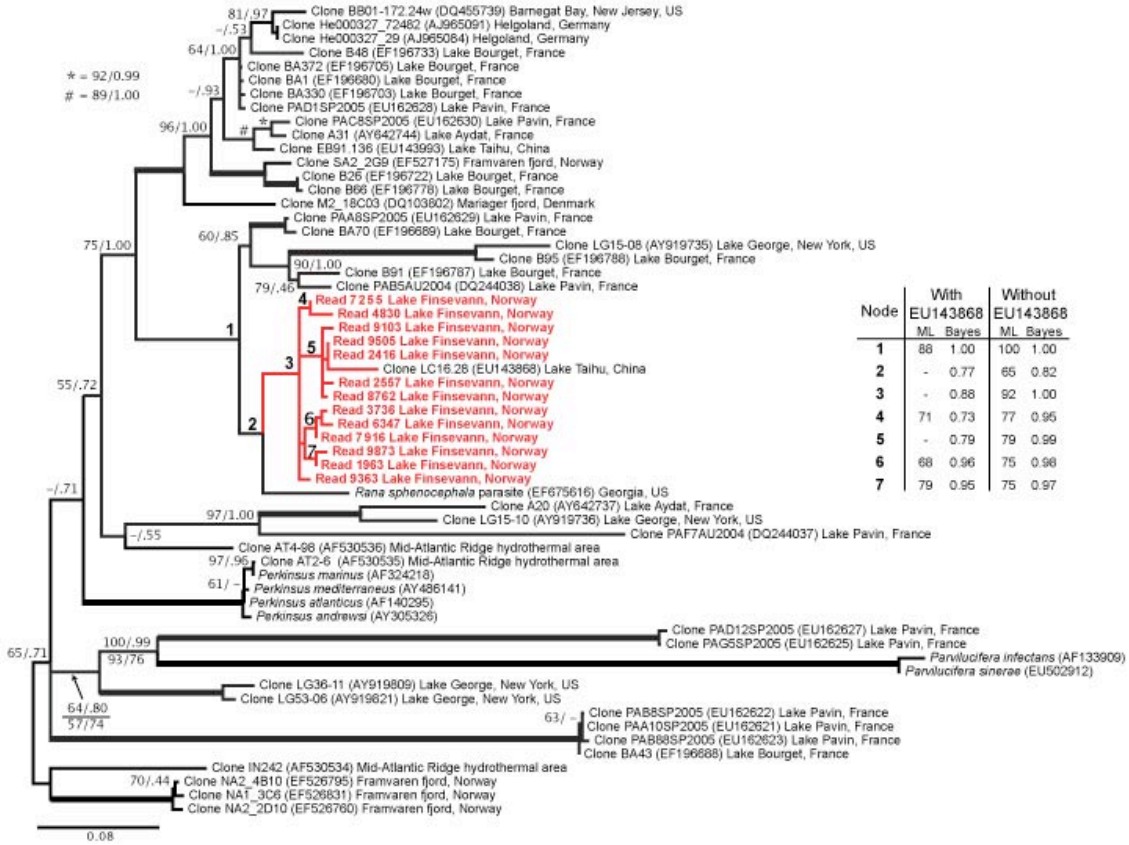


Figure 8. Bayesian phylogeny of Perkinsea based on an alignment consisting of 59 taxa and 1747 characters. Sequences generated from the Lake Finsevatn samples are shown in red Support values for the nodes affecting these reads are shown in the separate table, numbers 1-7 corresponds to the numbered nodes in the phylogenetic tree (modified from Paper 1 attached).

Similar to the alveolates, Haptophyta is one of the major lineages of eukaryotic algae and harbors many important primary producers. It is well described in marine environments worldwide, but only a few species have been described from freshwater environments (genera *Diacronema*, *Hymenomonas*, and *Chrysocromulina)*. Two of the reads from the pyrosequencing run 1were identified as haptophytes as shown in figure 9.

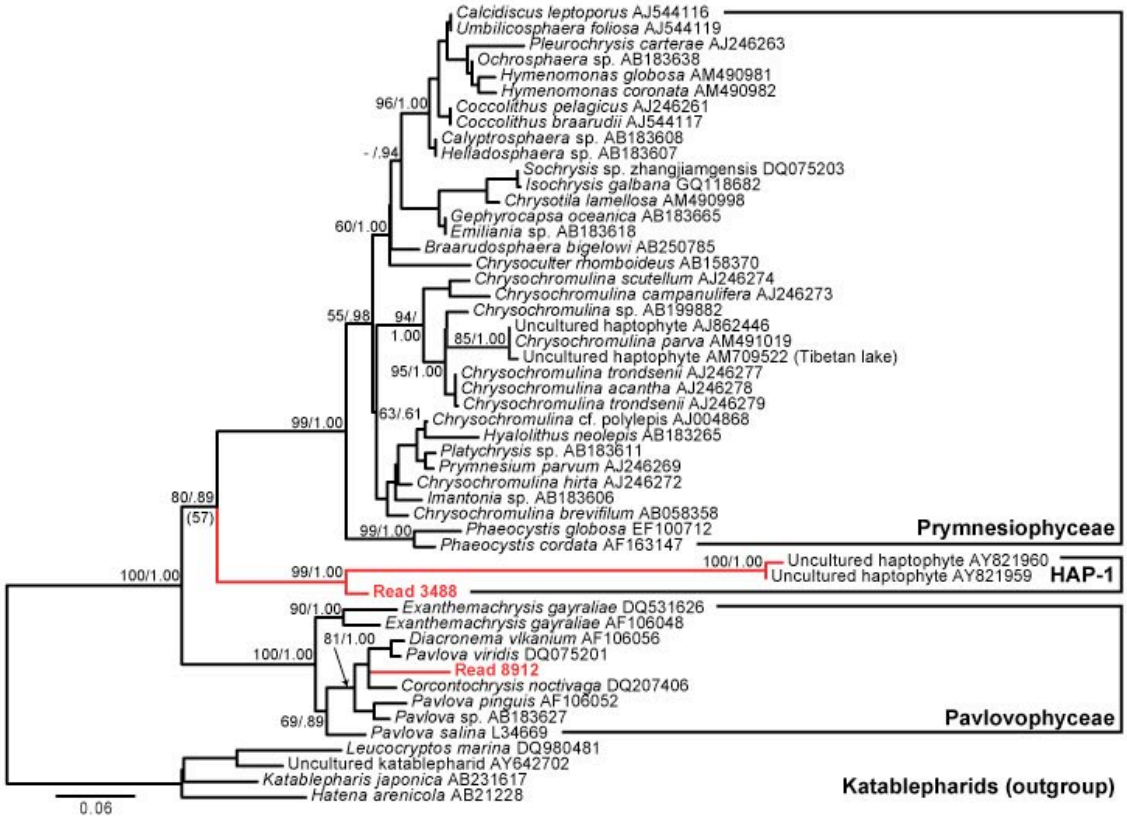


Figure 9. Maximum likelihood (ML) phylogeny of haptophytes derived from an 18S rDNA alignment consisting of 51 taxa and 1598 characters. Modified from figure 1 in Paper 2.

One sequence was placed firmly among the Pavlovophyceae, a group of haptophytes that has been observed in freshwater earlier, but observations have been limited to lakes with high salinity levels, springs and temporary ponds (Preisig 2002). The other sequence formed a sister group to Prymensiophyceae together with two environmental sequences from a French lake (accession numbers AY821960 and AY821959) (Slapeta et al 2005). These sequences have earlier been associated with Haptophyceae but with only low statistical support. The bootstrap support for this group is increased from 57% to 80% with the inclusion of our sequence. Hence, by applying our new primers optimized for 454 sequencing we could detect new haptophyte sequences that have only been observed once among all diversity surveys performed so far.

New group-specific primers reveals the presence of freshwater Pavlovophyceae

As we detected new sequences of haptophytes in freshwater, we wished to investigate this lineage more deeply by designing specific primers that only amplify the Pavlovophyeae lineage of the haptophytes. These primers were applied on DNA sampled from two other freshwater lakes nearby Oslo, Lake Sværsvann and Lake Pollen (Paper 2). Sequencing of cloned PCR products and preliminary phylogenetic analyses demonstrate that the haptophyte species detected in Finsevatn is also present in other types of lakes and has likely a much wider distribution.

Deeper sequencing gives largely the same overall diversity

With the discovery of such large diversity and many novel groups in the first pilot pyrosequencing run of Finsevatn DNA, we hoped that we would reveal even more novel groups if we sequenced more reads. In the second round of pyrosequencing I repeated the amplification of the V4 region of the 18S from the same DNA isolation as before, and sequenced 32547 reads with an average length of 297 bp, the total length distribution is displayed in figure 10 and show similar dual peaks as found in the length distribution of reads produced by the first pyrosequencing run (figure 6), except that the peak for the short reads was here at about 80 bp. After quality assessment and the removal of low quality data 14963 reads remained that resulted in 1370 clusters.

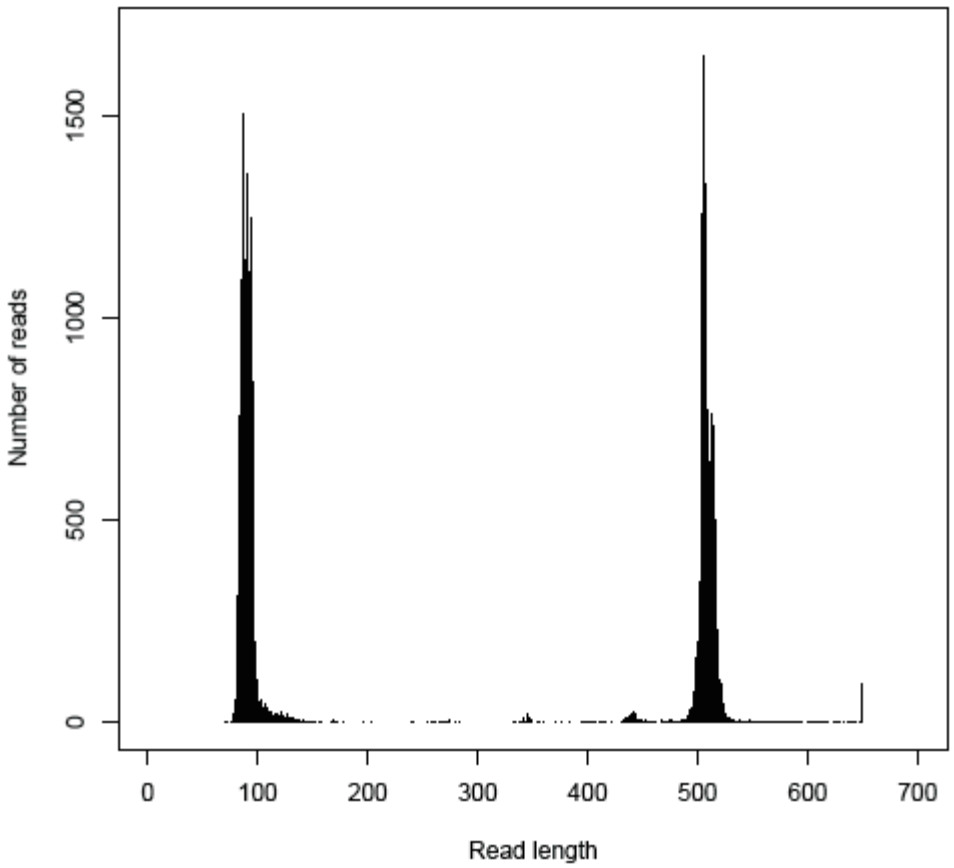


Figure 10. The length distribution of reads sequenced in the second round of pyrosequencing on the 454 GS FLX Titanium pyrosequencing platform.

This second round of pyrosequencing produced a diversity of protists very similar to that produced in the first round of sequencing, with all eukaryotic supergroups present except excavata (figure 11B). Since earlier microscopy investigations of Lake Finsevatn have repeatedly reported the presence of Euglenophyceae (Aanes et al 1997; Hagnar 2006), a group belonging to the excavates, this indicates that the designed primers are not suitable for these species, or that these Excavata species had a very low abundance at time of sampling. All other groups of protists described in Lake Finsevatn previously are recovered in both sequencing runs.

Large differences in abundance between the two sequencing runs

In contrast to the broad diversity pattern detected, there appears to be highly differences in the relative abundance of many of the major groups of eukaryotes that we sequenced. This is most clear for the cryptomonads, which had a much higher number of reads in the pilot sequencing than the second (and larger) sequencing run; i.e. 6.1% of the total clusters in run 1, but only 0.1% of the clusters in run 2 (figure 10). As the amplicons that are sequenced are amplified from the same DNA isolate, all differences in the abundance of sequenced reads must have been introduced in the PCR amplification.

PCR bias may be introduced into a generated dataset by two main processes during the PCR amplification: PCR selection and PCR drift (Wagner et al 1994). PCR selection is largely the result of different GC content at degenerate sites in the priming target areas, causing templates with a high GC content to be favored during amplification (Dutton et al 1993; Reysenbach et al 1992). PCR drift is caused by stochastic variation early in the amplification process and can cause different abundance of sequenced species in replicate PCR amplifications such as this (Polz and Cavanaugh 1998).

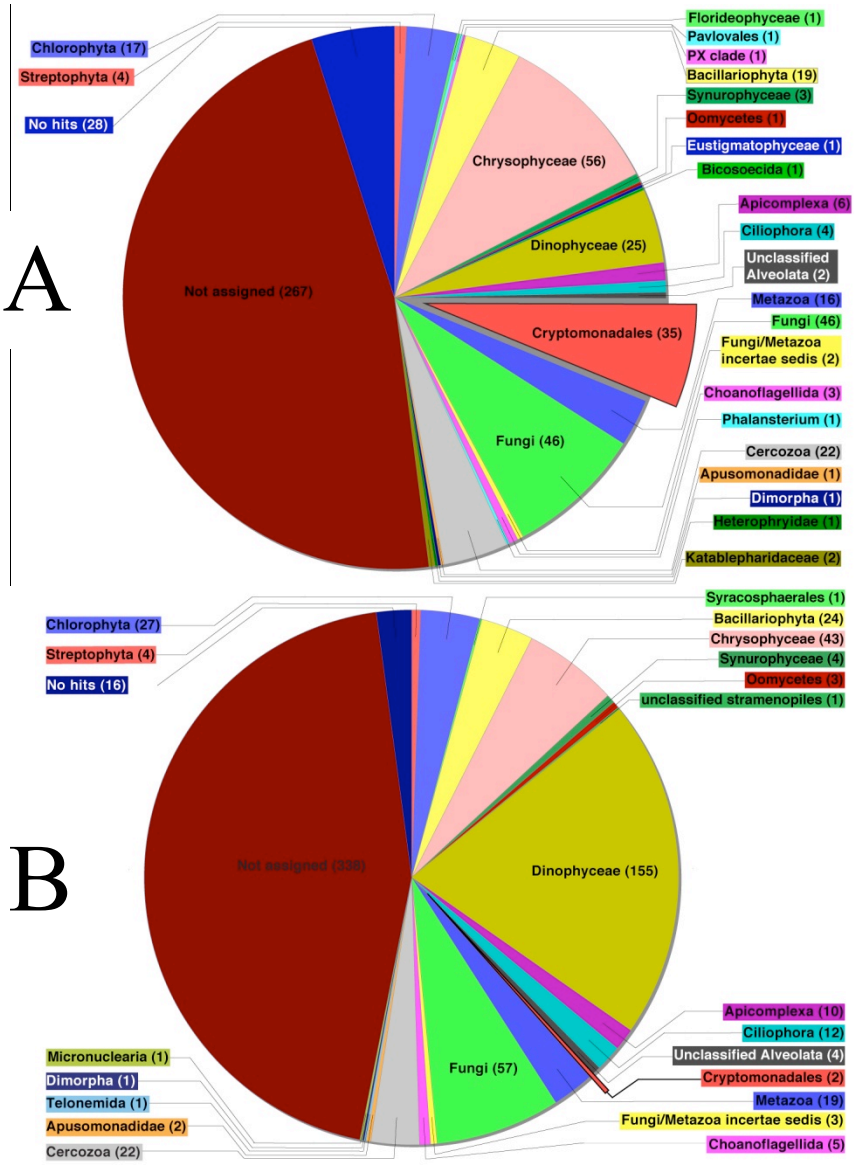


Figure 11. Displaying the diversity and abundance of the protists groups sequenced by the two pyrosequencing reactions. The numbers in parentheses indicate the number of clusters assigned to that protist group by the BLAST query parced with MEGAN. The group containing the cluster assigned as Cryptomonadales are pulled out as these sequences were extracted and used in a cryptomonad specific phylogeny.

A major question is whether the differences in abundance of reads between the two runs observed for many of the groups have substantial impact on the estimated species diversity. This question was addressed by using the cryptomonads as example. In order to evaluate the number of species of cryptomonads, the sequences assigned as cryptophytes were added to a cryptomonad alignment (Shalchian-Tabrizi et al 2008) together with the sequences extracted from the "not assigned" group in the MEGAN trees. Maximum likelihood analyses of this alignment showed that sequences from both sequencing runs were placed among typical

freshwater genera, such as CRY1, CRY2 (both identified by Shalchian-Tabrizi et al. 2008) and *Cryptomonas*. CRY1 and 2 are two clades almost entirely composed of environmental sequences and now also identified in sediments from Lake Finsevatn (figure 12). Hence at a genus level the PCR bias seem not to affect the taxon composition.

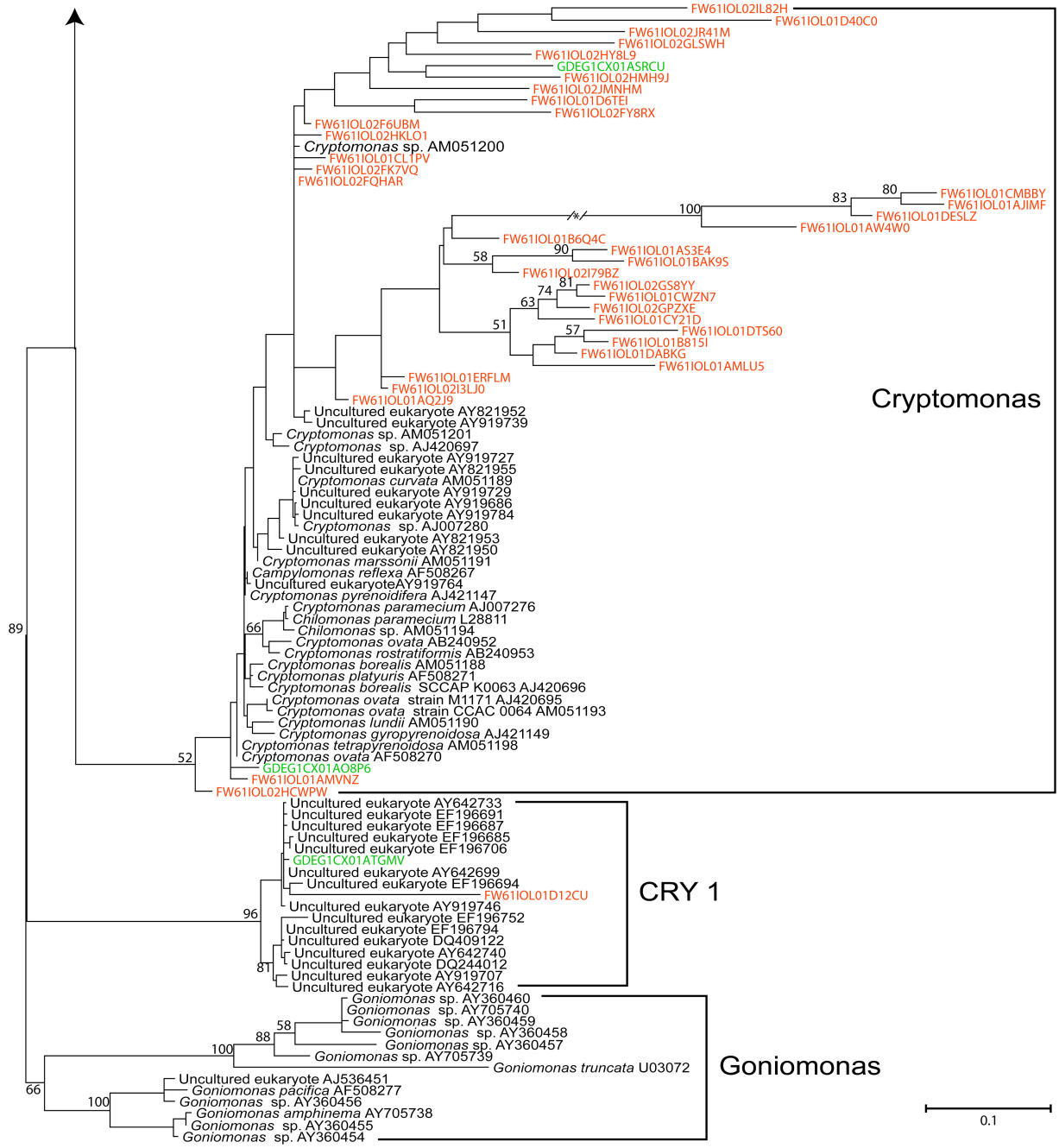


Figure 12. Maximum likelihood phylogeny of cryptomonad 18S rDNA sequences. Topology search and bootstrapping were done using the model GTR+G+I. Sequences in green are generated in pyrosequencing run 1 and sequences in red are generated in run 2.

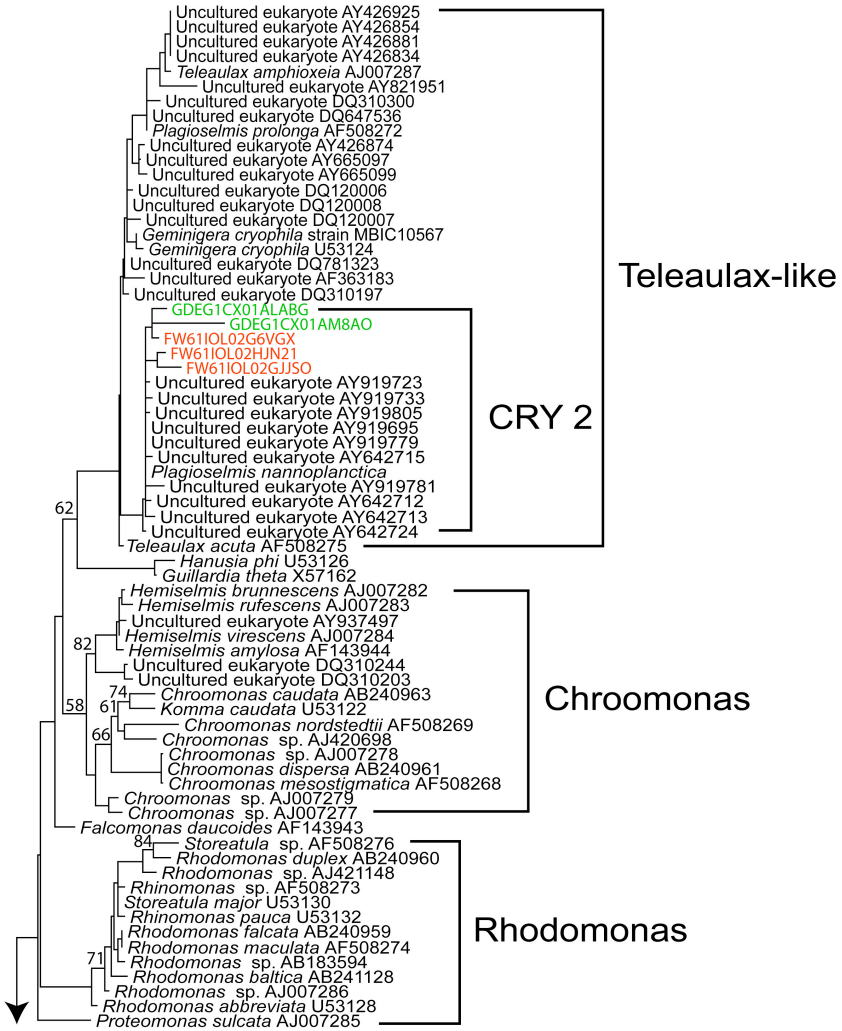


Figure 12 continued. See figure legend on previous page.

Conclusions and final remarks

The proposed protocol for 454 pyrosequencing of amplicons has proven successful. Large diversity of eukaryotes was uncovered in both the sequencing runs, covering all but one of the eukaryotic supergroups. More optimalization of the PCR protocol is needed in order to reduce the amount of primer dimer artifacts.

Novel groups not before observed in freshwater was also uncovered. This includes relatives if the marine parasites *Perkinsus* and *Parvilucifera* as well as the first truly freshwater sequence from the haptophytes group Pavlovophyceae.

A very important observation is that by increasing the sequencing effort threefold did not result in any significant increase in the revealed diversity. For the most common groups the level of diversity was about the same. As for the more rare groups, e.g. Telonemia and Apusomonas, they were not recovered in both runs indicating that several PCR amplifications and subsequent sequencing should be performed instead of deeper sequencing of a single PCR amplification.

A second very important observation was the fact that there were large differences in the abundance if the different groups generated by the two sequencing runs. As the two PCR amplifications were performed on the same DNA isolate, this difference in abundance must have been introduced by bias in the PCR. However, upon closer inscpection of a single group, the cryptomonads, we could see that there was not that many differences in the species detected, at least to a genus level. This shows that PCR based diversity studies should be used for quantitative measures with caution.

References

Amaral-Zettler L. A., McCliment E. A., Ducklow H. W., Huse S. M. 2009. A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes. *PLoS One* **4**(7):e6372.

Baas-Becking L. G. M. 1934. Geobiologie of Inleiding Tot de Milieukunde. Van Stockum & Zoon, The Hague.

Berney C., Fahrni J., Pawlowski J. 2004. How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biology* **2**:13.

Bråte J. 2008. Distribution and diversification processes in cryptomonads. Proalveolates and Telonemia investigated by environmental sequencing and Phylogenetic analyses. Master of science thesis. University of Oslo. Oslo.

Burki F., Shalchian-Tabrizi K., Minge M., Skjaeveland A., Nikolaev S. I., Jakobsen K. S. Pawlowski J. 2007. Phylogenomics reshuffles the eukaryotic supergroups. *PLoS One* **2**:e790.

Cavalier-Smith T., Lewis R., Chao E. E., Oates B., Bass D. 2009. Helkesimastix marina n. sp. (Cercozoa: Sainouroidea superfam. n.) a Gliding zooflagellate of novel ultrastructure and unusual ciliary behaviour. *Protist* **160**:452–479.

Davis A. K., Yabsley M. J., Keel M. K., Maerz J. C. 2007. Discovery of a novel alveolate pathogen affecting southern leopard frogs in Georgia: description of the disease and host effects. *Ecohealth* **4**:310–317.

DeLong E. F., Pace N. R. 2001. Environmental Diversity of Bacteria & Archaea. *Systematic Biology*. **50**:1-9.

Dutton C. M., Paynton C., Sommer S. 1993. General method for amplifying regions of very high G+C content. *Nucleic Acids Research.* **21**:2953–2954.

Engelbrektson A., Kunin V., Wrighton K. C., Zvenigorodsky N., Chen F., Ochman H., Hugenholtz P. 2010. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *ISMEJ* doi:10.1038/ismej.2009.153.

Hagnar R. 2006. Finsevatn og Flakavatn: En limnologisk undersøkelse. Master of science thesis. University of Oslo. Oslo.

Huson D. H., Auch A. F., Qi J., Schuster S. C. 2007 MEGAN Analysis of Metagenomic Data *Genome Research* **17**:377-386.

Kumar S., Carlsen T., Shalchian-Tabrizi K., Kauserud H. 2010. Phylosity – an online pipeline for processing 454 amplicon reads. In prep.

Lefèvre E., Roussel B., Amblard C., Sime-Ngando T. 2008. The Molecular Diversity of Freshwater Picoeukaryotes Reveals High Occurrence of Putative Parasitoids in the Plankton *PLoS One* **3**(6):e2324.

Lopez-Garcia P., Rodriguez-Valera F., Pedros-Alio C., Moreira, D. 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409:**603-607.

Loy A., Arnold R., Tischler P., Rattei T., Wagner M., Horn M. 2008. ProbeCheck - a central resource for evaluating oligonucleotide probe coverage and specificity. *Environmental Microbiology.* **10**:2894-2896.

Maddison W. P., Maddison D. R. 2009. Mesquite: a modular system for evolutionary analysis. Version 2.72 [http://mesquiteproject.org]

Marguiles M., Egholm M., Altman W. E., Attiya S., Bader J. S., Bemben L. A., Berka J., Braverman M. S., Chen Y. J., Chen Z., Dewell S. B., Lei D., Fierro J. M., Gomes X. V. Godwin B. C., He W., Helgesen S., Ho C. H., Irzyk G. P., Jando S. C., Alenquer M. L. I., Jarvie T. P., Jirage K. B., Kim J. B., Knight J. R., Lanza J. R., Leamon J. H., Lefkowitz S. M., Lei M., Li J., Lohman K. L., Lu H., Makhijani V. B., McDade K. E., McKenna M. P., Myers E. W., Nickerson E., Nobile J. R., Plant R., Puc B. P., Ronan M. T., Roth G. T., Sarkis G. J., Simons J. F., Simpson J. W., Srinivasan M., Tartaro K. R., Tomaz A., Vogt K. A., Volkmer S. H., Wang Y., Weiner M. P., Yu P., Begley R. F., Rothberg J. M.2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**:376-380.

McCarthy B. J., Bolton E. T. 1963. An approach to the measurements of genetic relatedness among organisms. *Proceedings of the National Academy of Sciences of the United States of America*. **50**:156-164.

Moon-van der Staay S. Y., De Watcher R., Vaulot D. 2001 Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**:607-610.

Nickrent D. L., Sargent M. L. 1991. An overview of the secondary structure of the V4 region of eukaryotic small-subunit ribosomal RNA. *Nucleic acids research* **19**(2):227-235.

Pace N. R., Stahl D. A., Lane D., Olsen G. J. 1986. The analysis of natural microbial populations by ribosomal RNA sequences. *Advances in Microbial Ecology* **9**:1–55.

Polz M. F., Cavanaugh C. M. 1998. Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology* **64:**3724–3730.

Posada D., Crandall K.A. 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* **14** (9):817-818.

Preisig H. R. 2002. Phylum Haptophyta (Prymnesiophyta). *University Press*: Cambridge.

Reysenbach A. L., Giver L. J., Wickham G. S., Pace N. R. 1992. Differential amplification of rRNA genes by polymerase chain reaction. *Applied and Environmental Microbiology* **58**:3417–3418.

Richards T. A., Vepritskiy A. A., Gouliamova D. E., Nierzwicki-Bauer S. A. 2005. The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages. *Environmental Microbiology* **7**(9):1413-1425.

Rozen S., Skaletsky H. J. 2000. Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics Methods and Protocols: Methods in Molecular Biology* 132:365-386.

Shalchian-Tabrizi K., Bråte J., Logares R., Klaveness D., Berney C., Jakobsen K. S. 2008. Diversification of unicellular eukaryotes: cryptomonad colonizations of marine and fresh waters inferred from revised 18S rRNA phylogeny. *Environmental Microbiology.* **10**:2635–2644.

Slapeta J., Moreira D., Lopez-Garcia P. 2005. The extent of protist diversity: insights from molecular ecology of freshwater eukaryotes. *Proceedings of the Royal Society Biological Sciences Series B* **272**(1576):2073-2081.

Stamatakis A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**:2688–2690.

Wagner A., Blackstone N., Cartwright P., Dick M., Misof B., Snow P., Wagner G. P., Bartels J., Murtha M., Pendleton J. 1994. Surveys of gene families using polymerase chain reaction: PCR selection and PCR drift. *Systematic Biology.* **43**:250–261.

Wheeler T., Kececioglu J. 2007. Multiple alignment by aligning alignments. *Bioinformatics* 23:i559-i568.

de Wit R., Bouvier T. 2006. 'Everything is everywhere, but, the environment selects'; what did Baas Becking and Beijerinck really say? *Environmental Microbiology* **8**(4):755-758.

Aanes K., Lien L., Brettum P. 1987. Resipientsituasjonen i Finsevatn 1985. Vurdering av behovet for rensetiltak. *NIVA rapport* O-85130.

Appendices

Paper 1

The ISME Journal (2010), 1–10

www.nature.com/ismej

ORIGINAL ARTICLE

Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental rDNA

Jon Bråte[1], Ramiro Logares[2], Cédric Berney[3], Dan Kristofer Ree[1], Dag Klaveness[1], Kjetill S. Jakobsen[1,4], Kamran Shalchian-Tabrizi[1]

[1]*Department of Biology, Microbial Evolution Research Group, University of Oslo, Oslo, Norway;* [2]*Department of Limnology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden;* [3]*Department of Zoology, University of Oxford, Oxford, UK and* [4]*Department of Biology, Centre for Ecological and Evolutionary Synthesis, University of Oslo, Oslo, Norway*

Protist parasites are ecologically important, as they can have great impact on host population dynamics and functioning of entire ecosystems. Nevertheless, little is known about their prevalence in aquatic habitats. Here, we investigate the diversity and distributional patterns of the protist parasites *Perkinsus* and *Parvilucifera* (Perkinsea). Our approach included 454 pyrosequencing of the 18S rDNA gene obtained from a high-altitude lake (Lake Finsevatn, Norway) and phylogenetic analyses of all publicly available sequences related to Perkinsea. The applied PCR primers target a 450 bp region that encompass the variable V4 region of the 18S rDNA gene and have been optimized for the Titanium upgrade of the 454 technology. Nearly 5000 sequences longer than 150 bp were recovered from nearly all eukaryotic supergroups, and of those, 13 unique sequences were affiliated to Perkinsea. Thus, our new strategy for 454 amplicon sequencing was able to recover a large diversity of distantly related eukaryotes and previously unknown species of Perkinsea. In addition, we identified 40 Perkinsea sequences in GenBank generated by other recent diversity surveys. Importantly, phylogenetic analyses of these sequences identified 17 habitat-specific marine and freshwater clades (PERK 1-17). Hence, only a few successful transitions between these habitats have taken place over the entire history of Perkinsea, suggesting that the boundary between marine and fresh waters may constitute a barrier to cross-colonizations for intracellular parasites.

The ISME Journal advance online publication, 15 April 2010; doi:10.1038/ismej.2010.39
Subject Category: Integrated genomics and post-genomics approaches in microbial ecology
Keywords: Perkinsea; *Perkinsus*; *Parvilucifera*; pyrosequencing; freshwater; sediments

Introduction

Parasitic unicellular eukaryotes (protists) can have considerable impact on ecosystem functioning. For example, the demise of large microalgal populations may affect the entire food chain in several aquatic ecosystems (Chambouvet *et al.*, 2008). The importance of parasites has been acknowledged in ecological and evolutionary studies on macroorganisms, however their impact on protistan host populations has so far been poorly investigated. New studies are suggesting the existence of a large diversity of protist parasites that infect other unicellular eukaryotes (Chambouvet *et al.*, 2008; Lefèvre *et al.*, 2008; Lepère *et al.*, 2008). Therefore, a major question is how these parasitic protists are distributed in spatiotemporal scales and whether they follow the distributions of their unicellular hosts.

The distribution of free-living microbes seems to be correlated with environmental salinity shifts. Recent studies indicate that saline and fresh waters contain phylogenetically distinct taxa, implying that cross-colonizations between these environments have been rare during the evolution of most eukaryotic and prokaryotic groups (reviewed in Logares *et al.*, 2009). As parasites usually need to be adapted to both the host and the extracellular environment, their spatiotemporal distribution patterns may therefore differ from that of free-living protists. In particular, parasites that spend most of their life cycle within the host's cell may be more prone to cross the saline–freshwater boundary, as their exposure to the extracellular aquatic environment would be minimal.

One of the most diverse groups of parasitic protists is Alveolata, in which Apicomplexa (for example, *Plasmodium*, *Cryptosporidium* and

Correspondence: K Shalchian-Tabrizi, Department of Biology, University of Oslo, Blindernveien 31, Oslo 0316, Norway.
E-mail: kamran@bio.uio.no
Received 3 December 2009; revised 12 February 2010; accepted 22 February 2010

Toxoplasma) and several lineages related to dinoflagellates (for example, Ellobiopsids, *Perkinsus* and *Parvilucifera*) seem to be entirely parasitic. In addition, a highly abundant and widespread group uncovered from environmental 18S rDNA libraries, the so-called marine alveolates (MA), appear to be largely parasitic (see Guillou *et al.*, 2008). To date, the *Perkinsus* and *Parvilucifera* (members of Perkinsea) and MA have only been identified in marine environments (Noren *et al.*, 1999; Villalba *et al.*, 2004; Groisillier *et al.*, 2006; Guillou *et al.*, 2008; Leander and Hoppenrath, 2008). However, microscopy observations, as well as reports of deeply diverging alveolate 18S rDNA sequences, point to the existence of freshwater Perkinsea species (Brugerolle, 2002, 2003; Green *et al.*, 2003; Richards *et al.*, 2005; Davis *et al.*, 2007; Lefèvre *et al.*, 2008; Lepère *et al.*, 2008).

Even though numerous surveys of the eukaryotic diversity have been undertaken in both marine and fresh waters over the last decade (López-García and Moreira, 2008), very few studies have focused on deeply diverging dinozoa other than the MA. Here, we investigate the diversity and distribution of another dinozoan group, Perkinsea, by searching publicly available sequence databases and by 454 pyrosequencing of 18S rDNA obtained from a high-mountain freshwater lake (Lake Finsevatn, Norway). We present a protocol for eukaryote-wide PCR amplification of the variable V4 region, optimized for Titanium upgrade of the GS FLX sequencing technology. Our results indicate a large and previously unknown diversity of Perkinsea, and provide rigorous evidence for the existence of species closely related to both *Perkinsus* and *Parvilucifera* in freshwater. In addition, our trees contained 17 new habitat-specific marine and freshwater clades (PERK 1-17), suggesting that cross-colonizations of marine and fresh waters have only taken place at a few occasions over the entire history of Perkinsea.

Materials and methods

Sample collection, DNA isolation and PCR amplification

Sediment samples were collected from Lake Finsevatn (60°36′ N—7°30 E) during March 2009. Lake Finsevatn is a high-mountain oligo- to mesotrophic lake situated at 1215 m above the sea level in an Arctic climate region. The samples were collected using a simple gravity corer at 18 m depth and the DNA was isolated from filtered cells using PowerSoil DNA isolation kit (MoBio, Carlsbad, CA, USA) following the manufacturers instructions. A PCR strategy aiming at amplifying the broadest possible eukaryotic diversity was designed for amplification and 454 pyrosequencing of the variable V4 region of the 18S rDNA gene. The length of the produced amplicon is about 450 bases and therefore suitable for the experimental conditions of the emulsion PCR. The amplification was carried out in two steps. The first by combining the universal forward primer 3NDf (Cavalier-Smith *et al.*, 2009) with the reverse primers V4_euk_R1 and V4_euk_R2 in two separate reactions. Adaptor A and B, as well as multiplexing tag (MID) were added to these amplicons in a second PCR by using template from the first PCR and composite primers (Table 1). The amplicons were pooled and cleaned on a Wizard SV column (Promega, Madison, WI, USA) before emulsion PCR. All amplifications were carried out in an Eppendorf Mastercycler ep (Eppendorf, Hamburg, Germany). Each amplification reaction (25 μl total) contained 7–50 ng of template DNA, 1 × DreamTaq Buffer with 2.5 mM Mg^{2+} (Fermentas, Burlington, Canada), 200 μM dNTPs, 0.2 μM of each primer and 0.6 U of DreamTaq DNA Polymerase (Fermentas). The PCR program for the first amplification round was: 94 °C for 2 min, followed by 34 cycles of 30 s at 94 °C, 30 s at 59 °C, 60 s at 72 °C with a final extension at 72 °C for 7 min. The second round of amplification was as follows: 94 °C for 2 min, followed by 15 cycles of 30 s at 94 °C, 30 s at 60 °C and 1 min at 72 °C, then 20 cycles of 30 s at 94 °C, 30 s at 65 °C and 1 min at 72 °C with a final extension for 7 min at 72 °C.

Pyrosequencing and removal of low-quality sequences

Amplicon pyrosequencing of the PCR products was carried out on a GS FLX Titanium machine (454 Life Sciences, Branford, CT, USA). We assessed sequence quality by using several criteria; reads that had degenerate bases, overall low-quality score and were shorter than 150 bp were removed. In addition, we

Table 1 Primer sequences for PCR and pyrosequencing used in this study

Primer name	*Sequence 5′–3′*	*Length (bp)*
3NDf	GGCAAGTCTGGTGCCAG	17
V4_euk_R1	GACTACGACGGTATCT(AG)ATC(AG)TCTTCG	27
V4_euk_R2	ACGGTATCT(AG)ATC(AG)TCTTCG	20
454_V4_euk_F1	**CCATCTCATCCCTGCGTGTCTCCGACTCAG***CGTGTCTCTA*AGGCAAGTCTGGTGCCAG	57
454_V4_euk_R1	**CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**GACTACGACGGTATCT(AG)ATC(AG)TCTTCG	57
454_V4_euk_R2	**CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**ACGGTATCT(AG)ATC(AG)TCTTCG	50

Composite primers are shown with pyrosequencing adaptors A and B (bold) and multiplex identifiers (italics).

removed identical reads (the longest reads were kept). All remaining sequences were BLASTed against the NCBI nr database (Altschul *et al.*, 1997). Sequences related to the alveolates were then selected and used in the phylogenetic analyses. Each of these alveolate sequences was carefully checked by manually inspecting the Phred quality scores to avoid inclusion of variable sites caused by sequencing artefacts. Indels in relation to polyhomomer regions usually had low Phred score, and such columns in the alignments were ignored in the phylogenetic analyses. In addition, apomorphic nucleotide characters may be generated by errors in the PCR and sequencing processes, producing an artificially high number of unique reads. Hence, for the phylogenetic analyses, we only approved Perkinsea sequences with variable characters that were shared by at least two independent sequences, of which the longest were used.

Data mining and identification of Perkinsea-related sequences

We identified publicly available environmental 18S rDNA sequences potentially related to Perkinsea by examining the results from several published surveys (Table S2). Sequences reported as either *Perkinsus* or likely to be related to *Perkinsus*, were used as queries in BLAST unrestricted and restricted (that is, using 'uncultured' and 'freshwater' as query limitations) searches against the NCBI nr data base (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi; Altschul *et al.*, 1997). BLAST searches were also carried out with *Perkinsus* sp. and *Parvilucifera* sp. as query sequences. To confirm that all new and downloaded sequences were evolutionarily related to Perkinsea, we analyzed them together with a broad sampling of eukaryotes (see Berney *et al.*, 2004). The alignment (AL1) consisted of 564 taxa and 1123 unambiguously aligned characters and was generated with MacClade v4.07 (Maddison and Maddison, 2000).

Alveolate and Perkinsea alignments

On the basis of the inferred global eukaryote tree from AL1, another alignment containing only alveolate taxa and the identified environmental sequences related to Perkinsea was constructed (AL2) to increase the number of unambiguously aligned characters (AL2 consisted of 130 taxa and 1443 nucleotide positions). The sequences were aligned with the MAFFT program online version 6.240 (http://align.bmr.kyushu-u.ac.jp/mafft/online/server/) (Katoh *et al.*, 2005), using the algorithm E-INS-I for an iterative refinement of the alignment that accounts for conserved motifs embedded among nonalignable regions. The alignment was manually edited and ambiguously aligned characters were excluded using MacClade v.4.07 (Maddison and Maddison, 2000). The tree produced from the alignment was used to identify and remove chimeric and essentially identical sequences (to reduce the alignment size), and subsequently the analyses were repeated (Figure 1). Three chimeric sequences were identified by visual inspection according to Berney *et al.* (2004). Two were removed (EU162624, EU162626) and the third (AY919735) was kept, but the chimeric section was removed (541 bp from the 3′-end). In addition, to further investigate the phylogenetic relationships within the Perkinsea, an alignment (AL3; 59 taxa and 1747 characters) containing only Perkinsea sequences was constructed based on AL 2. Furthermore, as all sequences generated in this study were covering the V4 region of the 18S rDNA gene, we assessed the impact of the short sequences on the tree topology by removing them in a separate phylogenetic analysis of the AL3 alignment (resulting alignment: 46 taxa and 1747 characters).

Phylogenetic analyses

Maximum likelihood analyses of all alignments were carried out using the program RaxML v.7.0.4 (Stamatakis, 2006). The general time reversible (GTR) model with parameters accounting for γ-distributed rate variation across sites (G) and invariable sites (I) was used in all analyses. Likelihood scores from 10 heuristic tree searches from random starting trees were generated, and the topology with the highest estimated likelihood was selected. Bootstrapping was carried out with 100 pseudoreplicates with the same evolutionary model as in the initial search, but with one heuristic search per replicate.

Bayesian phylogenies were reconstructed from alignments 2 and 3 using MrBayes v.3.0 (Ronquist and Huelsenbeck, 2003). In addition to the GTR + G + I model, the covarion parameters (that is, GTR + G + I + COV) was used to accommodate for different substitution rates across sequences. The four MCMC (Markov chain Monte Carlo) chains included three cold and one heated and they lasted for 4 000 000 generations. Two independent inferences, each from random starting trees, were performed. Posterior probabilities (pps) and mean marginal likelihood values of the trees were calculated after the burn-in phase, which was determined from the marginal likelihood scores of the initially sampled trees. The average split frequencies of the two runs was below 0.01, indicating convergence.

Removal of fast evolving taxa and characters with missing data

As *Parvilucifera* seem to be fast evolving, and hence may be prone to long-branch attraction artefacts, AL3 was analyzed both with and without *Parvilucifera* sequences, as well as associated long-branched taxa in the tree (EU162625 and

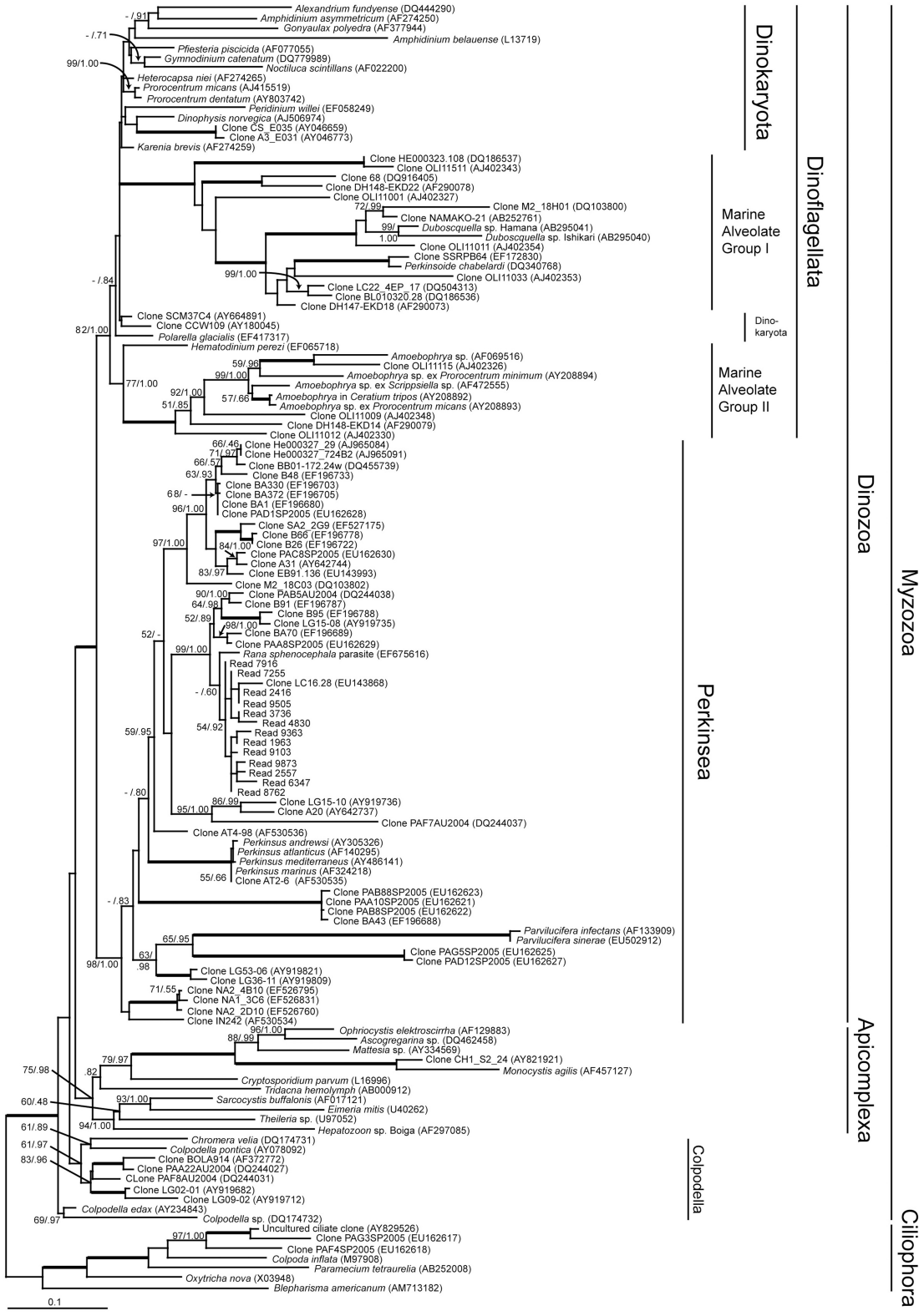


Figure 1 Bayesian phylogeny of alveolate 18S rDNA sequences reconstructed from an alignment consisting of 130 taxa 1443 characters with support values from maximum likelihood (ML) (GTR + G + I) and Bayesian (GTR + G + I + COV) inferences on the internal branches (ML bootstrap support (%)/Bayesian posterior probability values (pp)). Thick lines indicate support values of 1.00% and 1.00 pp. Values below 0.75 pp and 50% are not shown except for some backbone nodes.

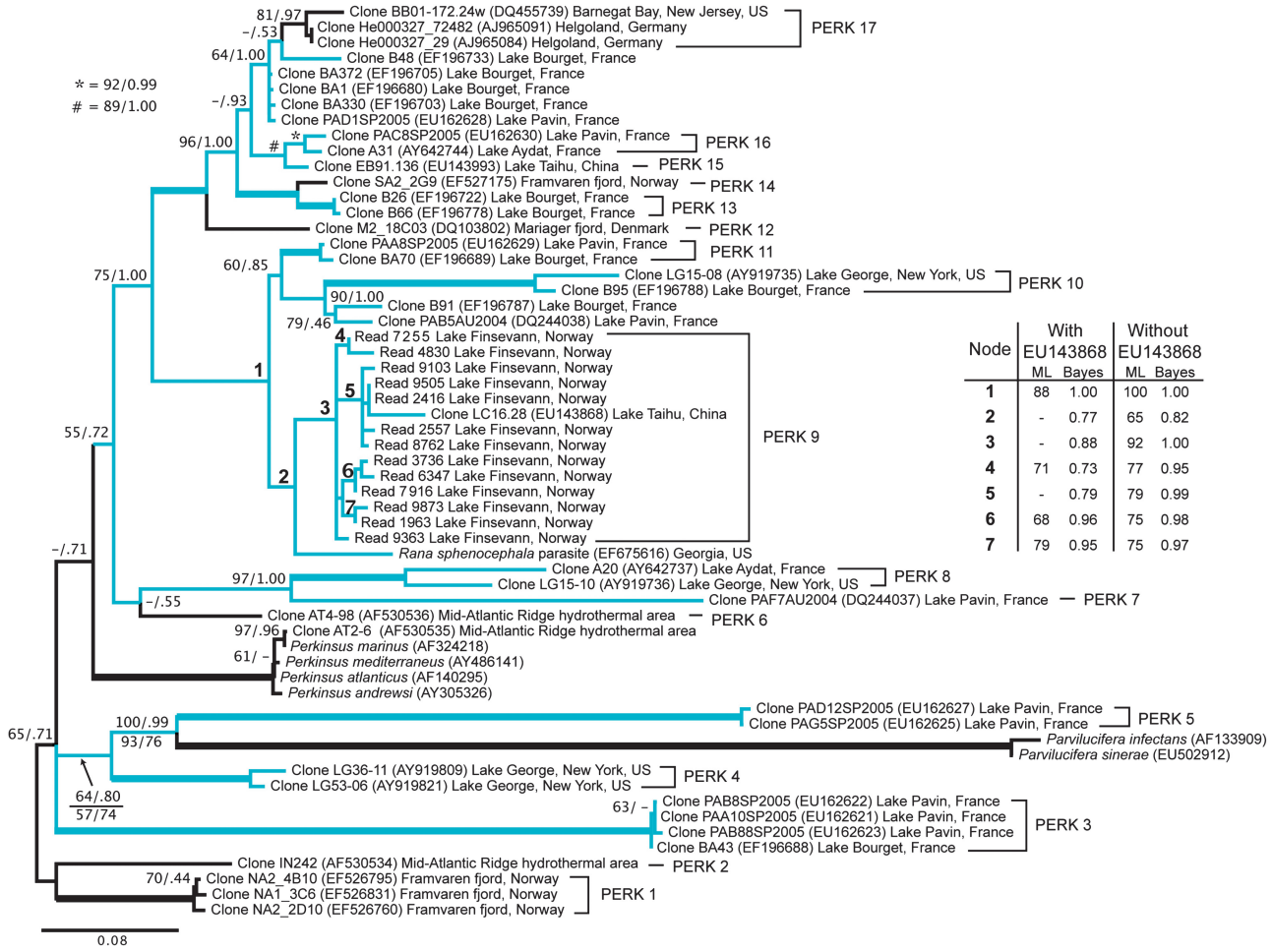


Figure 2 Bayesian phylogeny of Perkinsea 18S rDNA sequences inferred from an alignment consisting of 59 taxa and 1747 characters. See Figure 1 legend for description of analyses and support values. Blue lines indicate freshwater lineages and black lines indicate marine lineages. ML and Bayesian analyses were also performed without the sequence EU143868. Because of limited space, support values of the affected nodes are shown in a separate table with the corresponding nodes indicated with numbers at the nodes. The values at the nodes of *Parvilucifera* sp. and PERK 4 and 5 are showing the analyses with the fastest evolving sites removed: The values below the nodes represent ML analyses with category 8 and category 8 + 7 removed, respectively.

EU162627). To further test the impact of homoplasy and saturation, fast evolving sites were identified and removed using the AIR program package Kumar *et al.*, 2009). The site rates were calculated using the HKY85 substitution model and the tree topology obtained in the maximum likelihood analysis of alignment 3 described above. The site rates were divided into eight categories, with category 8 being the fastest evolving, and alignments were made by removing category 8 and 8 + 7. Removing category 8 left 1574 characters and removing category 8 + 7 left 1380 characters. All the phylogenetic analyses were performed on the freely available Bioportal at the University of Oslo (http://www.bioportal.uio.no).

To assess the impact of missing data in the environmental sequences, we deleted terminal regions of AL3 that contained missing nucleotides in several of the environmental sequences. As a consequence, some taxa were left with insufficient number of characters: EU143868, EU143993, EF196688 and EF196680. These were removed, resulting in an alignment consisting of 55 taxa and 435 characters.

Testing statistically the phylogenetic separation between marine and freshwater lineages

The statistical test of the separation between marine and freshwater lineages in the inferred Perkinsea clade (as presented in Figure 2) was carried out using the UniFrac program (Lozupone and Knight, 2005; http://bmf2.colorado.edu/unifrac/index.psp). The UniFrac significance test included 100 permutations, and the *P*-value was corrected for multiple comparisons using the Bonferroni's correction.

Results and discussion

New freshwater Perkinsea and improved phylogeny of enigmatic dinozoan lineages

The pyrosequencing of 18S rDNA V4 amplicons generated from Lake Finsevatn generated about

10 000 reads with an average length of 237 bp. Trimming of low-quality reads and reads shorter than 150 bp left 4769 sequences for further analyses. Of these, 53 reads were categorized as highly similar to Perkinsea on the basis of BLAST searches against the NCBInr database, resulting in 13 unique sequences after quality assessment. In addition to the new sequences, we identified 40 different 18S rDNA sequences of putative Perkinsea origin in publicly available databases, and thus, our study includes one of the largest sets of environmental sequences potentially related to *Perkinsus* and *Parvilucifera* that has been compiled to date.

Owing to the increased taxon selection in this branch of the alveolate diversity and by analyzing the sequences with the covarion evolutionary model, we obtained a high support for the clustering of *Perkinsus* and *Parvilucifera* together in the class Perkinsea (Cavalier-Smith, 2004) (Figure 1). The grouping of Perkinsea together with the MA and the dinoflagellates into the subphylum dinozoa received maximum support (100% maximum likelihood bootstrap support and 1.00 Bayesian pp), with the dinoflagellates being closest to the MA (82%, 1.00 pp). The branching order of these groups is highly congruent with recent classification of the alveolates (Cavalier-Smith, 2004), but whether the MA should be included among the dinoflagellates or as sisters to them is still unclear (see Massana *et al.*, 2008). This unresolved branching order between the dinoflagellates and the MA (Figure 1) could be due to a rapid radiation, early in the evolution of dinoflagellates (Saldarriaga *et al.*, 2004; Shalchian-Tabrizi *et al.*, 2006); multiple gene phylogenies are needed to help determine the relationships between these groups. The apicomplexans were retrieved as monophyletic (75%, 0.98 pp), whereas *Colpodella* sp. was polyphyletic and branched together with *Chromera velia* as well as freshwater environmental sequences at the base of the Myzozoa, similar to other 18S rDNA phylogenies (Leander *et al.*, 2003b; Moore *et al.*, 2008).

Perkinsea was highly supported (98%, 1.00 pp) and consisted of several environmental sequences from marine and freshwaters, as well as a sequence from an uncultured frog parasite and sequences from cultured *Perkinsus* and *Parvilucifera* species (Figure 1). A more thorough phylogenetic analysis of Perkinsea (Figure 2) showed that, in addition to the *Perkinsus* and *Parvilucifera* lineages, Perkinsea could be further subdivided into several highly supported (>95%, 1.00 pp) clades and solitary sequences, hereafter named PERK 1–17 (Figure 2). In addition, several larger assemblages were supported. *Parvilucifera* and PERK 1–5 had basal positions within Perkinsea, but their interrelationship was only weakly supported. PERK 7–17 formed a large assembly of uncultured sequences excluding *Perkinsus* and *Parvilucifera*. This assembly was almost exclusively composed of freshwater sequences and encompassed large sequence variation. Several solitary sequences, for example, PERK 6 and 12, are excluded from other clades with high support and could belong to larger, undersampled groups. The removal of the 13 sequences generated in this study did not change the tree topology in any significant way (see Supplementary Figure 1).

The taxonomic rank of the PERK groups revealed in the 18S rDNA phylogeny is not clear. However, the branch lengths of these sequences (that is, nucleotide variation), which are considerably longer than the marine *Perkinsus* sequences, indicates that each likely represent different species and that the different subgroups and solitary sequences constitute genera or higher order taxonomic levels.

Unknown Perkinsea diversity in marine and freshwaters

The genera *Perkinsus* and *Parvilucifera* have until now been regarded as strictly marine parasites (Delgado, 1999; Noren *et al.*, 1999; Erard-Le Denn *et al.*, 2000; Villalba *et al.*, 2004; Figueroa *et al.*, 2008; Leander and Hoppenrath, 2008). However, in recent environmental surveys of freshwater lakes, the existence of taxa related to Perkinsea has been suggested (Richards *et al.*, 2005; Lefèvre *et al.*, 2008; Lepère *et al.*, 2008). Nevertheless, as the analyses did not include any sequences from the MA or statistical support estimations for the tree topologies (Richards *et al.*, 2005; Lefèvre *et al.*, 2008; Lepère *et al.*, 2008), it cannot be determined whether the sequences belong to Perkinsea or to any of the MA.

Our extensive phylogenetic analyses of Perkinsea, including virtually all related environmental 18S sequences, showed a large and previously unknown diversity of both freshwater and marine taxa (Figure 2). Most of the subgroups within Perkinsea were entirely freshwater, except PERK 1, 2, 6, 12, 14 and 17, which were marine. The sequences from the Lake Finsevatn sediments formed the subgroup PERK 9 together with a sequence from a Chinese lake. However, as this sequence had very few overlapping characters with the sequences from this study, its position in the tree given in Figure 2 was very unstable. Analyses without this sequence substantially improved the support for several of the nodes while the branching order remained essentially identical (see Figure 2 legend). PERK 9 branched off together with the only environmental sequence of known origin, a sequence from a parasite of the leopard frog, *Rana sphenocephala* (EF675616). This fell into a large and diverse, exclusively freshwater group, including the subgroups PERK 9–11 (Figure 2). This parasite is known to cause massive killings of frogs by infecting the internal organs, especially the liver and kidneys (Green *et al.*, 2003; Davis *et al.*, 2007). There are also other reports of freshwater parasites with an outer morphology resembling *Perkinsus* (for example, a parasite of the freshwater cryptomonad *Chilomonas*

paramecium), but there are no sequences available for these species (Brugerolle, 2002, 2003). Considering the position of the *Rana* parasite among the Perkinsea sequences in the tree (Figure 2), as well as the evidence from other studies indicating that putative members of Perkinsea are infecting a wide range of species, such as mollusks, frogs, dinoflagellates and cryptomonads (Brugerolle, 2002, 2003; Green *et al.*, 2003; Villalba *et al.*, 2004; Davis *et al.*, 2007), we suspect that the entire Perkinsea is parasitic. This is very important to clarify because it would alter the present view of the role of picoeukaryotes in 'microbial loops' and the carbon cycle in aquatic systems. This view was recently presented by Lefèvre *et al.* (2008) as the 'parasite/saprotroph-dominated HF [heterotrophic flagellates] hypothesis' and should be further investigated by microscopy and fluorescence *in situ* hybridization methods.

In contrast to the MA, which have a worldwide distribution in the oceans, the Perkinsea seemed to predominate in fresh waters. Hence, each of these large putative parasitic groups seems to dominate different aquatic environments. This striking difference could potentially highlight important factors determining the dispersal and diversification of parasitic protists in aquatic environments. However, the sampling from different habitats and locations has not been equally extensive. In particular, freshwater habitats have so far been poorly sampled, and thus, more freshwater diversity may be revealed in future samplings. Furthermore, most of the environmental sequences related to Perkinsea have so far been obtained with universal PCR primers that often recover only a fraction of the total diversity. We therefore consider it likely that the total diversity of Perkinsea will be shown to be substantially larger by employing Perkinsea-specific primers in future environmental surveys.

Phylogeny of enigmatic perkinsoid species

The deep phylogenetic relationships within dinozoa have been difficult to resolve, particularly because of the rapid diversification and highly uneven evolutionary rates of many of the groups (Siddall *et al.*, 2001; Kuvardina *et al.*, 2002; Leander *et al.*, 2003a, 2003b; Cavalier-Smith and Chao, 2004; Silberman *et al.*, 2004; Groisillier *et al.*, 2006; Shalchian-Tabrizi *et al.* 2006). The increased taxon sampling of Perkinsea and application of covarion evolutionary models in this study has provided support for Perkinsea as sister to the dinoflagellates and MA. Nevertheless, the relationship between the two latter groups is still unresolved (Figure 1). It has been noted that the positions of *Perkinsus* and *Parvilucifera* in 18S rDNA phylogenies may be sensitive to the taxon sampling, and despite the increased taxon sampling here, long-branch artifacts could still affect their position (Cavalier-Smith and Chao, 2004). Therefore, to test the possibility that the position of *Parvilucifera* in our tree (Figure 2) was a result of long-branch attraction artifacts, the phylogenetic analyses were performed both with and without the two *Parvilucifera* sequences and the sequences EU162627 and EU162625. After removal of the two latter sequences, *Parvilucifera* still grouped together with AY919821 and AY919809, and the removal of the *Parvilucifera* sequences did not have any significant impact on the overall topology (results not shown). Removing fast evolving sites and missing terminal characters in the alignment did not change the position of *Parvilucifera*, although the bootstrap support was slightly altered (Figure 2).

Recently, two parasite species have been described as potential relatives of *Perkinsus*, a parasite of the southern leopard frog (*R. sphenocephala*) and a parasite of the Atlantic sardine (*Sardina pilchardus*), *Perkinsoide chabelardi* (Gestal *et al.*, 2006; Davis *et al.*, 2007). The *R. sphenocephala* parasite has been suggested to be a close relative of *Perkinsus* on the basis of morphology and phylogenetic analysis of the 18S rDNA gene (Davis *et al.*, 2007). Our trees confirm that the *R. sphenocephala* parasite belong to Perkinsea (Figure 2). In contrast, *P. chabelardi*, which has been previously determined as *Perkinsus*-like (Gestal *et al.*, 2006), is placed with high statistical support within group 1 of the MAs in our work (Figure 1), suggesting that previous classifications were incorrect (Gestal *et al.*, 2006). The placing of *P. chabelardi* in MA group 1, strengthens the view that this alveolate group is entirely parasitic (see Groisillier *et al.*, 2006; Harada *et al.*, 2006; Dolven *et al.*, 2007; Guillou *et al.*, 2008).

Marine–freshwater colonizations

In our phylogenetic reconstructions, the majority of the environmental sequences within Perkinsea fell into distinct clades according to their freshwater or marine origin (Figure 2). A UniFrac analysis of the Perkinsea clade, in which the fraction of unique branch lengths to each community against the total branch lengths between communities is measured based on a phylogenetic tree (see Lozupone and Knight, 2005), showed that this separation is statistically significant (P-value <0.01). This suggests that only a handful of marine–freshwater cross-colonization events have occurred during the evolution of Perkinsea. A few marine sequences are placed robustly within the freshwater clades, showing that recolonizations into the marine habitat have taken place as well (Figure 2). However, five sequences originated from fjords (EF526795, EF527175, EF526760, EF526831 and DQ103802) (Anke Behnke, personal communication) and could be the result of freshwater runoff. If this is indeed the case, PERK 1 would be entirely freshwater, and the large clade comprising PERK 12–17 would include only a single marine clade; hence, the number of colonization events would be reduced

considerably, and increase the possibility that Perkinsea originated in a freshwater habitat. Two marine sequences were sampled at the mid-Atlantic ridge (AF530536 and AF530534) and can hardly be the result of freshwater runoff. These two mid-Atlantic ridge sequences, together with the sequence DQ103802, have a solitary position within Perkinsea and could represent a larger, undersampled taxa hosting a greater diversity.

The limited transitions between marine and freshwater lineages are in great contrast to the wide geographic distribution of species within each type of habitat. Among the freshwater subgroups, sequences from as different locations as France, China and the United States are grouped together showing that dispersal between freshwater habitats over long distances has taken place without recolonizing the marine habitat.

Overall, our results on the putative parasites of Perkinsea show only a few successful cross-colonization events between marine and fresh waters during the evolutionary diversification of this group, implying that the biogeochemical differences between marine and fresh waters represent a strong barrier against cross-colonizations, concordant with recent results from other groups of free-living protists (Logares *et al.*, 2009).

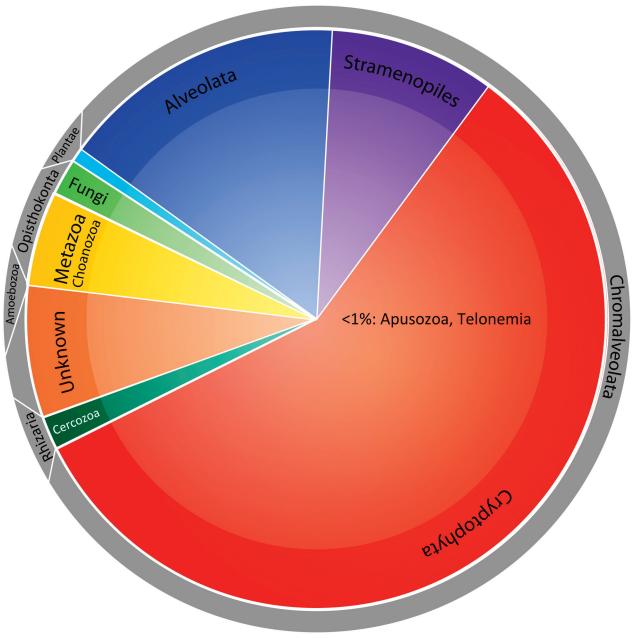


Figure 3 Pie chart showing the phylogenetic association of 4769 reads longer than 150 bp, as determined by BLAST searches against the NCBI nr database. Abundance of groups: cryptophyta, 57.50%; stramenopiles, 9.26%; alveolata, 15.92%; plantae, 0.77%; cercozoa, 1.80%; amoebozoa, 0.11%; fungi, 1.99%; choanozoa, 0.09%; metazoa, 5.21%; apusozoa, 0.02%; telonemia, 0.04%; unknown 7.29%.

Test of new 18S primers and pyrosequencing

Although our main goal with pyrosequencing of the lake Finsevatn sediments was to identify Perkinsea species, we could also uncover a large diversity of other eukaryote groups. In fact, all eukaryotic supergroups except Excavata could be clearly identified in our library (for description of supergroups, see Burki *et al.*, 2008), as well as several sequences highly similar to enigmatic lineages that traditionally have not been placed in any of these supergroups, such as Apusozoa and Telonemia (Figure 3). Altogether, the data show that the protocol presented here for PCR of the variable V4 18S rDNA region and 454 pyrosequencing is sensitive enough to detect even minute cell numbers, and hence are suitable for surveys of the overall eukaryote diversity. However, it should be noted that the abundance of the different groups was variable and could indicate that some groups are preferentially amplified over others; the chromalveolates are by far the most abundant group in the library, whereas only less than 1% belong to Amoebozoa.

Recently, the V9 region, which is typically about 150 bp long, was suggested as a suitable marker for diversity surveys of eukaryotes, and methods optimized for earlier generations of the 454 pyrosequencing technology has been used with success (Amaral-Zettler *et al.*, 2009; Stoeck *et al.*, 2009). As the PCR-based methods can be biased, it is likely that both the V4 and V9 regions could recover somewhat different diversity. It is a question whether the V4 or V9 region is most suitable for aberrant 18S genes, as often found in Foraminifera and other fast evolving species. A major advantage of the longer V4 region, which can now be covered with the Titanium upgrade, is the longer sequence lengths that allow for a more reliable taxon identification and more rigorous phylogenetic analyses. The many distantly related groups revealed in the data are likely to be real because of the substantial difference between the sequences that characterize the eukaryotic supergroups; however, because the quality assessment of sequences produced by the 454 and Sanger differs, it has been shown that the 454 data can overestimate the actual diversity in the environment (Quince *et al.*, 2009).

Acknowledgements

We thank the Bioportal team for implementing bioinformatics applications on the Bioportal (http://www.bioportal.uio.no). This work has been supported by the University of Oslo grants to KST, PhD fellowship to JB and travel funds to RL. The Swedish Research Council for Environment, Agricultural Sciences and Spatial Planning (FORMAS) via Uppsala Microbiomics Centre (UMC; http://www.microbiomics.se) has given financial support to RL.

References

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–3402.

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS ONE* **4**: e6372.

Behnke A, Bunge J, Barger K, Breiner H-W, Alla V, Stoeck T. (2006). Microeukaryote community patterns along an O2/H2S gradient in a supersulfidic anoxic fjord (Framvaren, Norway). *Appl Environ Microbiol* **72**: 3626–3636.

Berney C, Fahrni J, Pawlowski J. (2004). How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biol* **2**: 13.

Brugerolle G. (2002). *Cryptophagus subtilis*: a new parasite of cryptophytes affiliated with the Perkinsozoa lineage. *Eur J Protistol* **37**: 379–390.

Brugerolle G. (2003). Apicomplexan parasite Cryptophagus renamed Rastrimonas gen. nov. *Eur J Protistol* **39**: 101–101.

Burki F, Shalchian-Tabrizi K, Pawlowski J. (2008). Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes. *Biol Lett* **4**: 366–369.

Cavalier-Smith T. (2004). Chromalveolate diversity and cell megaevolution: interplay of membranes, genomes and cytoskeleton. In: Hirt RP, Horner DS (eds). *Organelles Genomes and Eukaryotic Phylogeny*. CRC Press: New York, pp 75–108.

Cavalier-Smith T, Chao EE. (2004). Protalveolate phylogeny and systematics and the origins of Sporozoa and dinoflagellates (phylum Myzozoa nom. nov.). *Eur J Protistol* **40**: 185–212.

Cavalier-Smith T, Lewis R, Chao EE, Oates B, Bass D. (2009). Helkesimastix marina n. sp. (Cercozoa: Sainouroidea superfam. n.) a Gliding zooflagellate of novel ultrastructure and unusual ciliary behaviour. *Protist* **160**: 452–479.

Chambouvet A, Morin P, Marie D, Guillou L. (2008). Control of toxic marine dinoflagellate blooms by serial parasitic killers. *Science* **322**: 1254–1257.

Davis AK, Yabsley MJ, Keel MK, Maerz JC. (2007). Discovery of a novel alveolate pathogen affecting southern leopard frogs in Georgia: description of the disease and host effects. *Ecohealth* **4**: 310–317.

Delgado M. (1999). A new 'diablillo parasite' in the toxic dinoflagellate Alexandrium catenella as a possibility to control harmful algal blooms. *Harmful Algae News* **19**: 1–3.

Dolven JK, Lindqvist C, Albert VA, Bjorklund KR, Yuasa T, Takahashi O *et al.* (2007). Molecular diversity of alveolates associated with neritic North Atlantic radiolarians. *Protist* **158**: 65–76.

Erard-Le Denn E, ChrÈtiennot-Dinet MJ, Probert I. (2000). First report of parasitism on the toxic dinoflagellate Alexandrium minutum Halim. *Estuar Coast Shelf Sci* **50**: 109–113.

Figueroa RI, Garcés E, Massana R, Camp J. (2008). Description, host-specificity, and strain selectivity of the dinoflagellate parasite Parvilucifera sinerae sp. nov. (Perkinsozoa). *Protist* **159**: 563–578.

Gestal C, Novoa B, Posada D, Figueras A, Azevedo C. (2006). *Perkinsoide chabelardi* n. gen., a protozoan parasite with an intermediate evolutionary position: possible cause of the decrease of sardine fisheries? *Environ Microbiol* **8**: 1105–1114.

Green DE, Feldman SH, Wimsatt J. (2003). Emergence of a Perkinsus-like agent in anuran liver during die-offs of local populations: PCR detection and phylogenetic characterization. *Proc Am Assoc Zoo Vet*, 120–121.

Groisillier A, Massana R, Valentin K, Vaulotl D, Guilloul L. (2006). Genetic diversity and habitats of two enigmatic marine alveolate lineages. *Aquat Microb Ecol* **42**: 277–291.

Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R *et al.* (2008). Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environ Microbiol* **10**: 3349–3365.

Harada A, Ohtsuka S, Horiguchi T. (2006). Species of the parasitic genus duboscquella are members of the enigmatic marine alveolate group I. *Protist* **158**: 337–347.

Katoh K, Kuma K-I, Toh H, Miyata T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucl Acids Res* **33**: 511–518.

Kumar S, Skjaeveland A, Orr R, Enger P, Ruden T, Mevik B-H *et al.* (2009). AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. *BMC Bioinformatics* **10**: 357.

Kuvardina ON, Leander BS, Aleshin VV, Myl'nikov AP, Keeling PJ, Simdyanov TG. (2002). The phylogeny of colpodellids (Alveolata) using small subunit rRNA gene sequences suggests they are the free-living sister group to apicomplexans. *J Eukaryot Microbiol* **49**: 498–504.

Leander BS, Clopton RE, Keeling PJ. (2003a). Phylogeny of gregarines (Apicomplexa) as inferred from small-subunit rDNA and beta-tubulin. *Int J Syst Evol Microbiol* **53**: 345–354.

Leander BS, Kuvardina ON, Aleshin VV, Mylnikov AP, Keeling PJ. (2003b). Molecular phylogeny and surface morphology of Colpodella edax (Alveolata): insights into the phagotrophic ancestry of apicomplexans. *J Eukaryot Microbiol* **50**: 334–340.

Leander BS, Hoppenrath M. (2008). Ultrastructure of a novel tube-forming, intracellular parasite of dinoflagellates: Parvilucifera prorocentri sp. nov. (Alveolata, Myzozoa). *Eur J Protistol* **44**: 55–70.

Lefèvre E, Roussel B, Amblard C, Sime-Ngando T. (2008). The molecular diversity of freshwater picoeukaryotes reveals high occurrence of putative parasitoids in the plankton. *PLoS ONE* **3**: e2324.

Lepère C, Domaizon I, Debroas D. (2008). Unexpected importance of potential parasites in the composition of the freshwater small-eukaryote community. *Appl Environ Microbiol* **74**: 2940–2949.

Logares R, Bråte J, Bertilsson S, Clasen JL, Shalchian-Tabrizi K, Rengefors K. (2009). Infrequent marine-freshwater transitions in the microbial world. *Trends Microbiol* **17**: 414–422.

López-García P, Moreira D. (2008). Tracking microbial biodiversity through molecular and genomic ecology. *Res Microbiol* **159**: 67–73.

Lozupone C, Knight R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228–8235.

Maddison D, Maddison W. (2000). *MacClade 4: Analysis of Phylogeny and Character Evolution*, 4 edn. Sinauer Associates: Sunderland, MA.

Massana R, Karniol B, Pommier T, Bodaker I, Oded B. (2008). Metagenomic retrieval of a ribosomal DNA repeat array from an uncultured marine alveolate. *Environ Microbiol* **10**: 1335–1343.

Moore RB, Obornik M, Janouskovec J, Chrudimsky T, Vancova M, Green DH *et al.* (2008). A photosynthetic alveolate closely related to apicomplexan parasites. *Nature* **451**: 959–963.

Noren F, Moestrup O, Rehnstam-Holm AS. (1999). Parvilucifera infectans Noren et Moestrup gen. et sp nov (Perkinsozoa phylum nov.): a parasitic flagellate capable of killing toxic microalgae. *Eur J Protistol* **35**: 233–254.

Quince C, Lanzen A, Curtis TP, Davenport RJ, Hall N, Head IM *et al.* (2009). Accurate determination of microbial diversity from 454 pyrosequencing data. *Nat Methods* **6**: 639–641.

Richards TA, Vepritskiy AA, Gouliamova DE, Nierzwicki-Bauer SA. (2005). The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages. *Environ Microbiol* **7**: 1413–1425.

Ronquist F, Huelsenbeck JP. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Saldarriaga JF, Max' Taylor FJR, Cavalier-Smith T, Menden-Deuer S, Keeling PJPJ. (2004). Molecular data and the evolutionary history of dinoflagellates. *Eur J Protistol* **40**: 85–111.

Shalchian-Tabrizi K, Minge MA, Cavalier-Smith T, Nedreklepp JM, Klaveness D, Jakobsen KS (2006). Combined Heat Shock Protein 90 and Ribosomal RNA Sequence Phylogeny Supports Multiple Replacements of Dinoflagellate Plastids. *J Eukaryot Microbiol* **53**: 217–224.

Siddall ME, Reece KS, Nerad TA, Burreson EM. (2001). Molecular determination of the phylogenetic position of a species in the genus colpodella (Alveolata). *Am Mus Novit* **3314**: 1–12.

Silberman JD, Collins AG, Gershwin LA, Johnson PJ, Roger AJ. (2004). Ellobiopsids of the genus Thalassomyces are alveolates. *J Eukaryot Microbiol* **51**: 246–252.

Stamatakis A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Stoeck T, Behnke A, Christen R, Amaral-Zettler L, Rodriguez-Mora M, Chistoserdov A *et al.* (2009). Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biology* **7**: 72.

Villalba A, Reece KS, Ordás MC, Casas SM, Figueras A. (2004). Perkinsosis in molluscs: a review. *Aquat Living Resour* **17**: 411–432.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)

Paper 2

biology letters

Freshwater haptophytes and ancient marine-freshwater colonization revealed by pyrosequencing of 18S rDNA

Journal:	*Biology Letters*
Manuscript ID:	Draft
Article Type:	Research
Date Submitted by the Author:	
Complete List of Authors:	Shalchian-Tabrizi, Kamran; University of Oslo, Department of Biology Røberg, Kjetil; University of Oslo, Department of Biology Ree, Dan; University of Oslo, Department of Biology Klaveness, Dag; University of Oslo, Department of Biology Bråte, Jon; University of Oslo, Department of Biology
Subject:	Evolution < BIOLOGY, Taxonomy and Systematics < BIOLOGY, Ecology < BIOLOGY, Molecular Biology < BIOLOGY, Bioinformatics < BIOLOGY
Categories:	Evolutionary Biology
Keywords:	Haptophytes, Pyrosequencing, Protists, Freshwater, Metagenomics, 18S

Freshwater haptophytes and ancient marine-freshwater colonization revealed by pyrosequencing of 18S rDNA

Kamran Shalchian-Tabrizi*, Kjetil Reier-Røberg*, Dan Kristofer Ree, Dag Klaveness, Jon Bråte^

Microbial Evolution Research Group, Department of Biology, University of Oslo, 1066 Blindern, 0316 Oslo, Norway

* Contributed equally to this work

^ Corresponding author

Short title: Freshwater haptophytes

Corresponding author:

Jon Bråte

Email: jon.brate@bio.uio.no, Phone: 0047 22855083

1. INTRODUCTION

Most lineages of aquatic unicellular eukaryotes are represented in both marine and freshwater environments. Diversity surveys of fresh waters using 18S rDNA clone libraries have recently revealed a large unknown freshwater diversity among several unicellular algal lineages such as cryptomonads, katablepharids and dinoflagellates (Logares et al 2007, Richards et al 2005, Shalchian-Tabrizi et al 2008, Slapeta et al 2006). In contrast, one of the most abundant and ecologically important groups of algae in marine ecosystems, the haptophytes, seems to be almost absent from freshwater.

Only a handful of haptophyte species have been observed visually in freshwater, including species of the genera *Diacronema*, *Hymenomonas*, *Chrysocromulina* and *Pavlova* with *Chrysochromulina parva* being by far the most dominant species in freshwater (Preisig 2002). DNA sequence information of typical freshwater haptophytes is even more limited, and is so far only available from *C. parva.*

Sequence data from environmental samples are somewhat more abundant. Two sequences from 18S rDNA clone libraries have been proposed to belong to haptophytes (Slapeta et al 2005). These sequences are apparently distantly related to the already sequenced *C. parva,* and are therefore possible examples of other freshwater haptophytes. However, these sequences are not firmly placed among the haptophytes in molecular phylogenies, and it is therefore not clear whether they are actually a sister to either of the two main classes of haptophytes, the Pavlovophyceae and Prymnesiophyceae (Cavalier-Smith et al 1996), or if these new environmental sequences represents a so far unknown sister group of haptophytes (Slapeta et al 2005). The position of these sequences in the

haptophyte phylogeny has implications on how often haptophytes have colonized freshwater.

In an earlier diversity survey of freshwater eukaryotes we investigated sediments from a high mountain lake in Norway, Lake Finsevatn. The target was the variable V4 region of the 18S rDNA gene and the method was optimized for emulsion PCR and sequencing with the Titanium upgrade of the 454 pyrosequencing technology (Bråte et al 2009). Two phylotypes belonging to the haptophytes were identified, each belonging to different and distantly related groups. The results demonstrate that haptophytes have colonized freshwater on at least three occasions and that the diversity in freshwater is likely to be larger than so far known.

2. MATERIALS AND METHODS

Sample collection, DNA isolation and amplicon generation

Sediment samples were collected from Lake Finsevatn (60˚36' N – 7˚30 E) on March 5th 2009, with a simple gravity corer at 18 m depth and filtered through a Millipore Durapore 0.22 µm filter (Millipore, MA, USA). DNA was isolated from the filter using the PowerSoil DNA isolation kit (MoBio, CA, USA) following the manufacturers instructions. PCR primers matching the variable V4 region of the 18S rDNA gene with the length of about 450bp was applied. The amplification was done in two steps. First, the forward primer 3NDf (5'- GGCAAGTCTGGTGCCAG-3') (Cavalier-Smith et al 2009) was combined with the reverse primers V4_euk_R1 (5'-GACTACGACGGTATCT(AG)ATC(AG)TCTTCG-3') and V4_euk_R2 (5'-

ACGGTATCT(AG)ATC(AG)TCTTCG-3') in two separate reactions. Second, primers for emulsion PCR (emPCR), sequencing and multiplexing tags (MIDs) were added to these initial amplicons using composite primers described in table 1. All amplifications were done on an Eppendorf Mastercycler ep (Eppendorf, Germany) with a sample volume of 25µl containing 7 – 50 ng of template DNA, 1x DreamTaq Buffer with 2.5mM Mg^{2+} (Fermentas, USA), 200µM dNTPs, 0.2µM of each primer and 0.6U DreamTaq DNA Polymerase (Fermentas, USA). The PCR program for the first amplification round was: 94°C for 2 min, followed by 35 cycles of 30 s at 94°C, 30 s at 59°C, 60 s at 72°C with a final extension at 72°C for 7 min. The second round of amplification was as follows: 94˚C for 2 min, followed by 15 cycles of 30 s at 94˚C, 30 s at 60˚C and 1 min at 72˚C, then 20 cycles of 30 s at 94˚C, 30 s at 65˚C and 1 min at 72 ˚C with a final extension for 7 min at 72˚C.

Pyrosequencing and quality assessment of sequences

Pyrosequencing of the amplicons were done by standard procedures and sequenced on a GS FLX Titanium machine following the manufacturers instructions (454 Life Sciences, CT, USA). Quality assessment was undertaken as by follows: 1) sequences shorter than 150 bp was removed, 2) sequences containing degenerate code N and/or with overall low quality scores were removed, 3) identical sequences were filtered out. All remaining sequences were BLASTed against the NCBInr database, and sequences potentially of haptophyte origin selected for phylogenetic analyses. The quality of these reads was further assessed by careful inspection of the phred quality scores to avoid inclusion of variable sites caused by sequencing artefacts. Indels in relation to polyhomomer regions usually had low phred score and were ignored in the phylogenetic analyses.

Alignments and phylogenetic analyses

Six reads were identified as potential haptophytes and were added to an 18S rDNA alignment containing a wide diversity of eukaryotes (see taxon sampling in Berney et al 2004; 190 taxa and 1286 characters). Two sequences were included among the haptophytes with high support (supplementary figure S1) and were added to an alignment of haptophyte sequences with katablepharids as outgroup, allowing for more unambiguously aligned positions (51 taxa and 1598 characters). Alignments are submitted as supplementary material.

Maximum likelihood trees were inferred using RAxML (Stamatakis 2006), using the general time reversible (GTR) model with gamma distributed (G) site-rate variation and a proportion of invariable sites (I). The gamma distribution was approximated using four rate categories and the most likely topology among 100 independent inferences was selected and bootstrapping was done with 100 pseudoreplicates.

Bayesian analyses were done with MrBayes 3.1.2 (Ronquist and Huelsenbeck 2003) using the model GTR+G+I. Prior settings were as default and the MCMC chains lasted for 4,000,000 generations, and trees were saved every 100th generations. After burn-in, which was based on visual inspection of the of the MCMC chains, the stationary phase was used for calculating the majority rule consensus tree and posterior probability values. Convergence of the MCMC chains was tested running the Bayesian inference twice from different random starting trees. When tree topologies, mean likelihood scores, and the posterior probability values showed almost identical results after burn-in from the independent runs, we assumed that the MCMC had lasted long enough to converge. In

addition, the corresponding covarion models were tested in MrBayes. The analyses using the model with covarion parameters yielded likelihood values significantly lower than the corresponding model without covarion parameters. The covarion analyses were therefore excluded from further consideration.

3. Results

Identification of sequences related to haptophytes

Among all sequences produced from Lake Finsevann, six reads were identified as putative haptophytes by BLAST searches against the NCBInr database (http://blast.ncbi.nlm.nih.gov/Blast.cgi). To confirm the evolutionary relationship to haptophytes, we reconstructed a maximum likelihood tree with all major groups of eukaryotes included (Supplementary Fig. 1). In this tree, which is congruent with other published 18S rDNA trees (Berney et al 2004, Shalchian-Tabrizi et al 2006), two of the new sequences fell within the haptophytes whereas four others were clearly not related to haptophytes and were omitted from further analyses.

A haptophyte phylogeny was constructed with only katablepharids as outgroup in order to include more unambiguously aligned sites in our alignment. The tree show that the two new haptophyte freshwater lineages detected from DNA sampled in Lake Finsevatn is placed within the haptophytes with high statistical support in both maximum likelihood and Bayesian inferences (100%,1.00pp, Fig. 1).

Each sequence formed distinct freshwater lineages of haptophytes (Fig. 1). The sequence, read 8912, was firmly placed among the Pavlovophyceae, while the second sequence, read 3488, did not belong to either Pavlovophceae or Prymnesiophyceae, but branched off as sister to Prymensiophyceae together with two previously published environmental sequences (accession numbers AY821960 and AY821959); because the group is distinct and excluded from both of the haptophytes classes, the group is here named HAP-1. Furthermore, both sequences were statistically supported to constitute

freshwater lineages distantly related to the group comprising *Crysochromulina parva* and other related freshwater sequences.

4. Discussion

Haptophytes colonized freshwater more than once

The presented data clearly demonstrate a hitherto unknown diversity of haptophytes in freshwater, and since these freshwater species are distantly related, each may represent independent marine to freshwater transitions. The 18S rDNA phylogeny clearly suggests that none of these transitions are related to the colonization that has taken place among the *Chrysochromulina*. Species of Pavlovophyceae have been observed in freshwater, especially *Diakronema vlkanium*, but these are usually observed in saline lakes and springs and temporary ponds (Preisig 2002). Our sequence, read 8912, comes from the sediments of a truly freshwater lake and provides rigorous evidence that also Pavlovophyceae species exists in freshwater.

Interestingly, the other haptophytes sequence we generated (read 3488) is closely related to the two sequences recently obtained from a French lake (Slapeta et al 2005). These sequences were earlier associated with haptophytes but not firmly placed among them because of insufficient statistical support (Slapeta et al 2005). The phylogenetic reconstruction (Fig. 1) shows with relatively high bootstrap values and posterior probabilities that these two sequences, and the new sequence generated here, together constitute a distinct lineage (HAP-1).

The increased statistical support for placing HAP-1 as sister lineage to prymnesiophyceae is likely due to increased taxon sampling that reduces the long (basal) branch to the HAP-1 clade. Accordingly, when the sequence read 3488 was removed from the analysis, the bootstrap support for the placement of HAP-1 was significantly reduced from 80% to 57% (Fig. 1). It is therefore likely that the placement of HAP-1 will become even more supported as the long branches of the group is divided by addition of new sequences. If this topology is confirmed by future studies, it implies that the ancestor of the Prymnesiophyceae lineage diverged into separate marine and freshwater lineages. According to recent dating estimations the extant Prymnesiophyceae radiated about 540 MYA (Liu et al 2009). By now placing HAP-1 as the sister to the earlier known Prymnesiophyceae, the freshwater colonization may have taken place even earlier.

The morphology and biology of species belonging to the HAP-1 clade is unclear, as microscopic observations have not been linked to these sequences. Nevertheless, as freshwaters in general have been investigated thoroughly by visual inspections, it is tempting to speculate that these species may be very small (pico-eukaryotes) or are intracellular organisms; either parasites or symbionts. It is also possible that the species have been observed, but not recognised as haptophytes due to very aberrant morphology. Linking HAP-1 and environmental sequences to cells by FISH or similar methods is needed in order to better understand the diversity and biological significance of these species.

Haptophytes belong phylogenetically to a major group of eukaryotes that also contain cryptomonads, katablepharids and possibly also centrohelids and Telonemia (altogether named CCTH or Hacrobia; Burki et al 2009, Okamoto et al 2009). All these

groups have in earlier or current works showed a larger freshwater diversity than previously thought. With this new data from haptophytes, all lineages that constitute the CCTH assemblage have successfully colonized freshwater at more than one occasion. As shown here for the haptophytes, some of the transitions occurred early in almost all these lineages (see review in Logares et al 2009).

The data obtained with pyrosequencing show that there may be higher diversity among haptophytes that earlier thought. The newly proposed procedure (Bråte et al 2009) that combine massive parallel sequencing and the improvement of sequence read length allow the use of the V4 region in 18S rDNA for diversity studies. The data presented here demonstrate that applying massive parallel sequencing technology can detect even minute number of cells and hence applicable for molecular surveys of haptophyte diversity in lakes.

Acknowledgements

We thank the 454 sequencing lab and Bioportal computing facility at University of Oslo (UiO) for sequencing and computer resources. We thank Ramiro Logares for discussions and Cédric Berney for 18S rDNA alignments and assistance with chimera check of sequences. This work was financed by a starting grant to KST from UiO.

REFERENCES

Berney C, Fahrni J, Pawlowski J (2004). How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biol* **2:** 13.

Bråte J, Logares R, Berney C, Ree DK, Klaveness D, Jakobsen KS *et al* (2009). Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental DNA. *ISME Journal***:** Submitted.

Burki F, Inagaki Y, Brate J, Archibald JM, Keeling PJ, Cavalier-Smith T *et al* (2009). Large-Scale Phylogenomic Analyses Reveal That Two Enigmatic Protist Lineages, Telonemia and Centroheliozoa, Are Related to Photosynthetic Chromalveolates. *Genome Biol Evol* **2009:** 231-238.

Cavalier-Smith T, Allsopp MTEP, Häuber MM, Gothe G, Chao EE, Couch JA *et al* (1996). Chromobiote phylogeny: the enigmatic algae *Reticulosphaera japonensis* is an aberrant haptophyte, not a heterokont. *Eur J Phycol* **31:** 255-263.

Cavalier-Smith T, Lewis R, Chao EE, Oates B, Bass D (2009). Helkesimastix marina n. sp. (Cercozoa: Sainouroidea superfam. n.) a Gliding Zooflagellate of Novel Ultrastructure and Unusual Ciliary Behaviour. *Protist* **160:** 452-479.

Liu H, Aris-Brosou S, Probert I, de Vargas C (2009). A timeline of the environmental genetics of the haptophytes. *Mol Biol Evol***:** msp222.

Logares R, Shalchian-Tabrizi K, Boltovskoy A, Rengefors K (2007). Extensive dinoflagellate phylogenies indicate infrequent marine-freshwater transitions. *Mol Phylogenet Evol* **45:** 887-903.

Logares R, Bråte J, Bertilsson S, Clasen JL, Shalchian-Tabrizi K, Rengefors K (2009). Infrequent marine-freshwater transitions in the microbial world. *Trends Microbiol* **17:** 414-422.

Okamoto N, Chantangsi C, Horák A, Leander BS, Keeling PJ (2009). Molecular Phylogeny and Description of the Novel Katablepharid *Roombia truncata* gen. et sp. nov., and Establishment of the Hacrobia Taxon nov. *PLoS ONE* **4:** e7080.

Preisig HR (2002). *Phylum Haptophyta (Prymnesiophyta)*. University Press: Cambridge.

Richards TA, Vepritskiy AA, Gouliamova DE, Nierzwicki-Bauer SA (2005). The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages. *Environ Microbiol* **7:** 1413-1425.

Ronquist F, Huelsenbeck JP (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572-1574.

Shalchian-Tabrizi K, Eikrem W, Klaveness D, Vaulot D, Minge MA, Le Gall F *et al* (2006). Telonemia, a new protist phylum with affinity to chromist lineages. *Proc Biol Sci* **273:** 1833-1842.

Shalchian-Tabrizi K, Bråte J, Logares R, Klaveness D, Berney C, Jakobsen KS (2008). Diversification of unicellular eukaryotes: Cryptomonad colonisations of marine and fresh waters inferred from revised 18S rRNA phylogeny. *Environ Microbiol* **10:** 2635-2644.

Slapeta J, Moreira D, Lopez-Garcia P (2005). The extent of protist diversity: insights from molecular ecology of freshwater eukaryotes. *Proc R Soc Biol Sci Ser B* **272:** 2073-2081.

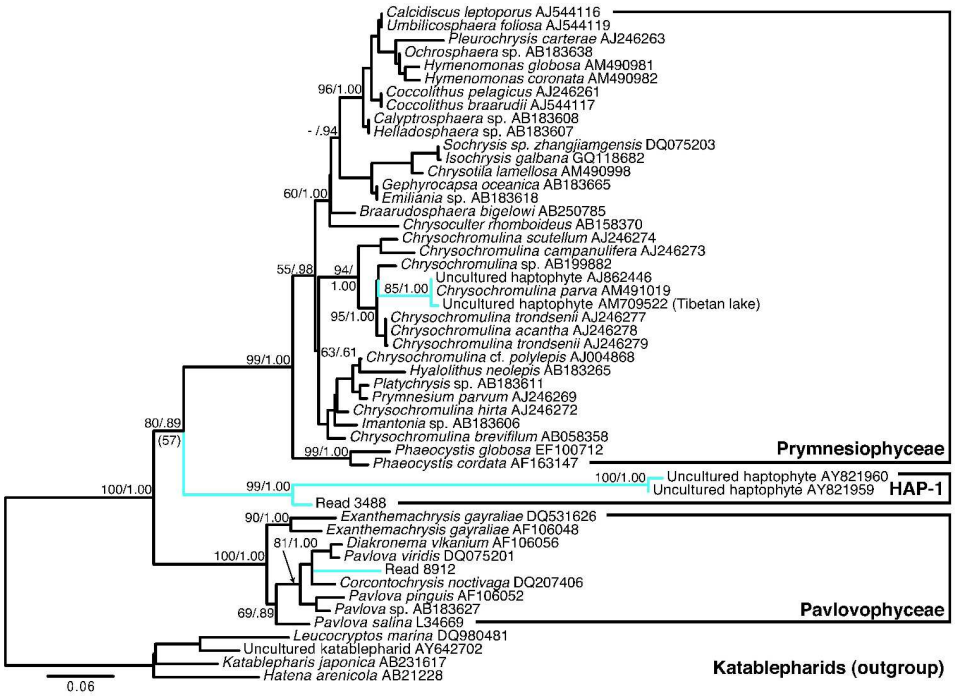
Slapeta J, Lopez-Garcia P, Moreira D (2006). Present status of the molecular ecology of kathablepharids. *Protist* **157:** 7-11.

Stamatakis A (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22:** 2688-2690.

Figure legends:

Figure 1: Maximum likelihood (ML) phylogeny of haptophytes derived from an 18S rDNA alignment consisting of 51 taxa and 1598 characters. Values at nodes correspond to ML and Bayesian analyses (ML/Bayes) using the model GTR+G+I. Value in brackets show the ML bootstrap value when the sequence read 3488 was removed. Only values on selected nodes are shown. Black lines indicate marine lineages while blue lines indicate freshwater lineages.

Table 1. Showing the composite primer sequences used in this study. Adaptor sequences are shown in bold, Multiplex Identifiers (MID) are shown in italics and primer sequences are in regular type.



Maximum likelihood (ML) phylogeny of haptophytes derived from an 18S rDNA alignment consisting of 51 taxa and 1598 characters. Values at nodes correspond to ML and Bayesian analyses (ML/Bayes) using the model GTR+G+I. Value in brackets show the ML bootstrap value when the sequence read 3488 was removed. Only values on selected nodes are shown. Black lines indicate marine lineages while blue lines indicate freshwater lineages.
268x188mm (600 x 600 DPI)

Primer name	Sequence 5' – 3'	Length (bp)
454_euk_F1	**CCATCTCATCCCTGCGTGTCTCCGACTCAG***CGTGTCTCTA*AGG CAAGTCTGGTGCCAG	57
454_euk_R1	**CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**GACTACGACGG TATCT(AG)ATC(AG)TCTTCG	57
454_euk_R2	**CCTATCCCCTGTGTGCCTTGGCAGTCTCAG**ACGGTATCT(AG) ATC(AG)TCTTCG	50